Lipschitz Certificates for Layered Network Structures Driven by Averaged Activation Operators

Patrick Combettes, Jean-Christophe Pesquet

Lipschitz Certificates for Layered Network Structures Driven by Averaged Activation Operators*

Patrick L. Combettes[1] and Jean-Christophe Pesquet[2]

[1]North Carolina State University, Department of Mathematics, Raleigh, NC 27695-8205, USA
plc@math.ncsu.edu

[2]CentraleSupélec, Inria, Université Paris-Saclay, Center for Visual Computing, 91190 Gif sur Yvette, France
jean-christophe@pesquet.eu

Abstract

Obtaining sharp Lipschitz constants for feed-forward neural networks is essential to assess their robustness in the face of perturbations of their inputs. We derive such constants in the context of a general layered network model involving compositions of nonexpansive averaged operators and affine operators. By exploiting this architecture, our analysis finely captures the interactions between the layers, yielding tighter Lipschitz constants than those resulting from the product of individual bounds for groups of layers. The proposed framework is shown to cover in particular many practical instances encountered in feed-forward neural networks. Our Lipschitz constant estimates are further improved in the case of structures employing scalar nonlinear functions, which include standard convolutional networks as special cases.

1 Introduction

Artificial neural networks are becoming increasingly central tools in tasks such as learning, modeling, data processing, and decision making. As first noted in [52], neural networks are vulnerable to adversarial examples which, though close to other data inputs, lead to very different outputs. This potential lack of stability makes the networks vulnerable and unreliable in key application areas; see, for instance, [1, 30, 35] and the references therein. To protect networks against such instabilities various techniques have been explored [39, 43, 44, 54]. Although these defense strategies may be effective in certain scenarios, they do not provide formal guarantees of robustness for general networks and they have been shown to be breakable by new attacks; see, for instance, [3, 18].

It has been acknowledged for some time that the Lipschitz behavior of a network plays a key role in the analysis of its robustness [52]. Simply put, if a layered network is modeled by an operator T acting between normed spaces, with Lipschitz constant θ, given an input x and a perturbation z, we can majorize the perturbation on the output via the inequality

$$\|T(x+z)-Tx\| \leqslant \theta\|z\|. \tag{1.1}$$

Thus θ can be used as a certificate of robustness of the network provided that it is tightly estimated. Lipschitz regularity is also an important ingredient in the derivation of generalization bounds and

*Contact author: P. L. Combettes, plc@math.ncsu.edu, phone: +1 919 515 2671. The work of P. L. Combettes was supported by the National Science Foundation under grant CCF-1715671. The work of J.-C. Pesquet was supported by Institut Universitaire de France.

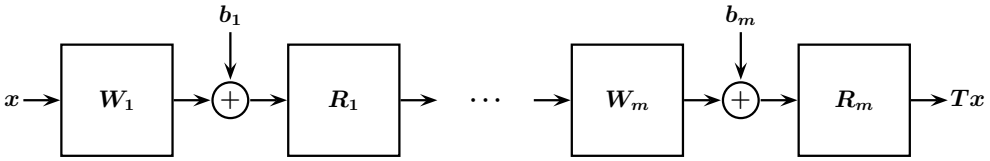

Figure 1: In Model 1.1, the ith layer involves a linear weight operator W_i, a bias vector b_i, and an activation operator R_i, which is assumed to be a nonlinear averaged nonexpansive operator.

approximation bounds [6, 11, 50], and of reachability conditions [47]. In [52] the estimation of θ is performed by evaluating the Lipschitz constant of the layers individually and then defining θ as the product of these constants, which typically yields pessimistic bounds. Lipschitz constants have also been computed for specific situations, e.g., [5, 33, 49, 53]. Overall, however, deriving analytically accurate constants for general contexts remains an open problem. The objective of the present paper is to address this question for a general class of layered networks. Mathematically, our network model is described as an alternation of affine and nonlinear operators. This type of structure also arises in variational and equilibrium problems, as well as in network science, e.g., [16, 24, 27, 56]. Adopting the same terminology as in the neural network literature, where they model the activity of neurons, the nonlinear operators will be called activation operators. Our stability analysis focuses on the following m-layer model, in which the activation operators are averaged nonexpansive operators (see Fig. 1). Recall that an operator $R\colon \mathcal{H} \to \mathcal{H}$ acting on a Hilbert space $\mathcal{H}$ is α-averaged for some $\alpha \in [0,1]$ if there exists a nonexpansive (i.e., 1-Lipschitzian) operator $Q\colon \mathcal{H} \to \mathcal{H}$ such that

$$R = (1-\alpha)\,\mathrm{Id} + \alpha Q. \tag{1.2}$$

In other words, $R = \mathrm{Id} + \alpha(Q - \mathrm{Id})$ is an underrelaxation of a nonexpansive operator (see [8] for a detailed account). This class of operators was introduced in [4] and shown in [21] to model various problems in nonlinear analysis as it includes common operators such as projection operators, proximity operators, resolvents of monotone operators, reflection operators, gradient step operators, and various combinations thereof. Recent theoretical developments and applications to data science include [9, 10, 12, 13, 15, 22, 26, 34, 41, 51, 55, 56].

Model 1.1 Let $m \geqslant 1$ be an integer and let $(\mathcal{H}_i)_{0\leqslant i\leqslant m}$ be nonzero real Hilbert spaces. For every $i \in \{1,\ldots,m\}$, let $W_i\colon \mathcal{H}_{i-1} \to \mathcal{H}_i$ be a bounded linear operator, let $b_i \in \mathcal{H}_i$, let $\alpha_i \in [0,1]$, and let $R_i\colon \mathcal{H}_i \to \mathcal{H}_i$ be an α_i-averaged operator. Set

$$T = T_m \circ \cdots \circ T_1, \quad \text{where} \quad (\forall i \in \{1,\ldots,m\}) \quad T_i\colon \mathcal{H}_{i-1} \to \mathcal{H}_i\colon x \mapsto R_i(W_i x + b_i). \tag{1.3}$$

Since the operators $(R_i)_{1\leqslant i\leqslant m}$ are nonexpansive, a Lipschitz constant for T in (1.3) is

$$\theta_m = \prod_{i=1}^{m} \|W_i\|. \tag{1.4}$$

However, as already mentioned, this constant is usually quite loose and of limited use to assess the actual stability of the network. A novelty of our approach is to take into account the averagedness properties of the individual activation operators to capture more sharply the overall interactions between the layers, yielding tighter constants than those provided by computing bounds for groups of layers. Our specific contributions are the following:

- We show that the most common activation operators used in neural networks are averaged operators. This not only provides an a posteriori justification for Model 1.1, but also indicates that this highly structured framework should be of interest in the analysis of other properties of layered networks beyond stability.
- We derive a general expression for a Lipschitz constant of T in terms of the averagedness constants of the activation operators $(R_i)_{1\leqslant i\leqslant m}$ and the norms of certain compositions of the linear operators $(W_i)_{1\leqslant i\leqslant m}$. This Lipschitz constant is shown to lie between the simple upper bound (1.4) and the lower bound $\|W_m\circ\cdots\circ W_1\|$ corresponding to a purely linear network. Our analysis applies to any type of linear operator, in particular convolutive ones, and it does not require any additional assumptions on the activation operator. In particular, differentiability is not assumed and our results therefore cover, in particular, networks using the rectified linear unit (ReLU) and max-pooling operations.
- In the common situation when the activation operators are separable, we obtain tighter Lipschitz constants for various norms.
- Under some positivity condition, we prove that a Lipschitz constant of the network reduces to that of the associated purely linear network obtained by removing the nonlinear operators.

In [24], we investigated the special case of Model 1.1 in which the activation operators $(R_i)_{1\leqslant i\leqslant m}$ are proximity operators, hence $1/2$-averaged (see Section 3.1). The objective there was to study the asymptotic behavior of deep network structures rather than their stability.

The remainder of the paper is organized as follows. In Section 2 we present an illustration of our main result in a simple special case. In Section 3.1 we provide the necessary nonlinear analysis background. In Section 3.2 we show that a wide array of activation operators used in neural networks are indeed nonexpansive. In Section 4 we derive general results concerning Lipschitz constants for Model 1.1. Section 5 refines this analysis in the case of separable activation operators.

2 Preview of the main results in a simple scenario

We illustrate on a simple instance the main results of the paper. More precisely, we consider a three-layer ($m=3$) network where, for every $i\in\{0,1,2,3\}$, $\mathcal{H}_i$ is the standard Euclidean space $\mathbb{R}^{N_i}$. In this case, each linear operator W_i is identified with a matrix in $\mathbb{R}^{N_i\times N_{i-1}}$. To further simplify our setting, we assume that the operators R_1, R_2, and R_3 correspond to ReLU layers, that is, for each $i\in\{1,2,3\}$,

$$\big(\forall x=(\xi_k)_{1\leqslant k\leqslant N_i}\in\mathbb{R}^{N_i}\big)\quad R_ix=\big(\rho(\xi_k)\big)_{1\leqslant k\leqslant N_i},\quad\text{where}\quad\rho\colon\xi\mapsto\max\{0,\xi\}. \tag{2.1}$$

In view of (1.2), $\rho=(1/2)\,\mathrm{Id}+(1/2)|\cdot|$ is $1/2$-averaged since $|\cdot|$ has Lipschitz constant 1. This implies that the operators R_1, R_2, and R_3 are also $1/2$-averaged [24]. Let us now introduce two parameters which will play a central role in our analysis, namely,

$$\theta_3=\frac{1}{4}\big(\|W_3W_2W_1\|+\|W_3W_2\|\,\|W_1\|+\|W_3\|\,\|W_2W_1\|+\|W_3\|\,\|W_2\|\,\|W_1\|\big) \tag{2.2}$$

and

$$\vartheta_3=\sup_{\Lambda_1\in\mathscr{D}_{\{0,1\}}^{(N_1)},\Lambda_2\in\mathscr{D}_{\{0,1\}}^{(N_2)}}\|W_3\Lambda_2W_2\Lambda_1W_1\|, \tag{2.3}$$

where $\|\cdot\|$ is the spectral norm and, for each $i \in \{1,2\}$, $\mathscr{D}_{\{0,1\}}^{(N_i)}$ denotes the set of $N_i \times N_i$ diagonal matrices with entries in $\{0,1\}$. In this context, our main result states that both θ_3 and ϑ_3 are Lipschitz constants of the network, and that

$$\|W_3W_2W_1\| \leqslant \vartheta_3 \leqslant \theta_3 \leqslant \|W_3\|\,\|W_2\|\,\|W_1\|. \tag{2.4}$$

In addition, if the entries of the matrices $(W_i)_{1\leqslant i\leqslant 3}$ are in $[0,+\infty[$, then a Lipschitz constant of the network is $\|W_3W_2W_1\|$.

Example 2.1 To illustrate the improvement of the proposed bound over the classical product norm estimate, we consider a fully connected network with $N_0 = 8$, $N_1 = 10$, $N_2 = 6$, and $N_3 = 3$. The entries of the matrices $(W_i)_{1\leqslant i\leqslant 3}$ are generated randomly and independently according to a normal distribution. We evaluate the Lipschitz constant estimate θ_3 provided by (2.2) and the lower bound in (2.4). The average (resp. minimal) value of $\theta_3/(\|W_1\|\,\|W_2\|\,\|W_3\|)$ computed over 1000 realizations is approximately equal to 0.6699 (resp. 0.5112), while the average (resp. minimal) value of $\|W_3W_2W_1\|/(\|W_1\|\,\|W_2\|\,\|W_3\|)$ is approximately equal to 0.3747 (resp. 0.1208). In addition, the average (resp. minimal) value of $\vartheta_3/(\|W_1\|\,\|W_2\|\,\|W_3\|)$ computed over 1000 realizations is approximately equal to 0.4528 (resp. 0.2424). In agreement with (2.4), this estimation of the Lipschitz constant is better than θ_3 and significantly sharper than $\|W_1\|\,\|W_2\|\,\|W_3\|$.

In the remainder of this paper, we show that the above results hold in a much more general context (for an arbitrary number of layers m, arbitrary Hilbert spaces, and a wide class of activation operators), and that some of them can be extended to non-Euclidean norms. To establish these results, we need to introduce suitable mathematical tools in the next section.

3 Nonexpansive averaged activation operators

3.1 Nonlinear analysis tools and notation

We review some key facts and definitions which will be used subsequently; see [8] for further information. Throughout, $\mathcal{H}$ is a real Hilbert space with power set $2^{\mathcal{H}}$, scalar product $\langle\cdot\mid\cdot\rangle$, and associated norm $\|\cdot\|$.

Let $R\colon \mathcal{H}\to\mathcal{H}$ be an operator and let $\alpha\in[0,1]$. Then R is nonexpansive if it is 1-Lipschitzian, α-averaged if there exists a nonexpansive operator $Q\colon\mathcal{H}\to\mathcal{H}$ such that $R = (1-\alpha)\,\mathrm{Id}+\alpha Q$, and firmly nonexpansive if it is $1/2$-averaged. Let $A\colon\mathcal{H}\to 2^{\mathcal{H}}$ be a set-valued operator. We denote by $\operatorname{gra} A = \big\{(x,u)\in\mathcal{H}\times\mathcal{H}\;\big|\;u\in Ax\big\}$ the graph of A and by A^{-1} the inverse of A, i.e., the operator with graph $\big\{(u,x)\in\mathcal{H}\times\mathcal{H}\;\big|\;u\in Ax\big\}$. In addition, A is monotone if

$$(\forall (x,u)\in\operatorname{gra}A)(\forall (y,v)\in\operatorname{gra}A)\quad \langle x-y\mid u-v\rangle\geqslant 0, \tag{3.1}$$

and maximally monotone if there exists no monotone operator $B\colon\mathcal{H}\to 2^{\mathcal{H}}$ such that $\operatorname{gra}A\subset\operatorname{gra}B\neq\operatorname{gra}A$. If A is maximally monotone, then its resolvent $J_A = (\mathrm{Id}+A)^{-1}$ is firmly nonexpansive. We denote by $\Gamma_0(\mathcal{H})$ the class of proper lower semicontinuous convex functions from $\mathcal{H}$ to $]-\infty,+\infty]$. Let $f\in\Gamma_0(\mathcal{H})$. The conjugate of f is

$$\Gamma_0(\mathcal{H})\ni f^*\colon u\mapsto \sup_{x\in\mathcal{H}}(\langle x\mid u\rangle - f(x)) \tag{3.2}$$

and the subdifferential of f is the maximally monotone operator

$$\partial f\colon\mathcal{H}\to 2^{\mathcal{H}}\colon x\mapsto\big\{u\in\mathcal{H}\;\big|\;(\forall y\in\mathcal{H})\;\;\langle y-x\mid u\rangle + f(x)\leqslant f(y)\big\}. \tag{3.3}$$

For every $x \in \mathcal{H}$, the unique minimizer of $f + \|x - \cdot\|^2/2$ is denoted by $\operatorname{prox}_f x$. We have $\operatorname{prox}_f = J_{\partial f}$ and prox_f is therefore firmly nonexpansive.

Let C be a nonempty convex subset of $\mathcal{H}$. Then ι_C is the indicator function of C (it takes values 0 on C and $+\infty$ on its complement) and $d_C\colon x \mapsto \min_{y\in C} \|x - y\|$ is its distance function. If C is closed, its projection operator is $\operatorname{proj}_C = \operatorname{prox}_{\iota_C}$.

3.2 Activators as averaged operators

We show via various illustrations that the assumption made in Model 1.1 on the activation operators covers many existing instances of feed-forward neural networks. Let us start with some key properties.

Proposition 3.1 *Let $\mathcal{H}$ be a real Hilbert space, let $\alpha \in [0,1]$, and let $R\colon \mathcal{H} \to \mathcal{H}$ be α-averaged. Then the following hold:*

(i) *There exist a maximally monotone operator $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ and a constant $\lambda \in [0,2]$ such that $R = \operatorname{Id} + \lambda(J_A - \operatorname{Id})$. Furthermore, if $\lambda \leqslant 1$, then R is firmly nonexpansive.*

(ii) *Suppose that $\mathcal{H} = \mathbb{R}$. Then there exist a function $\phi \in \Gamma_0(\mathbb{R})$ and a constant $\lambda \in [0,2]$ such that $R = \operatorname{Id} + \lambda(\operatorname{prox}_\phi - \operatorname{Id})$. Furthermore, R is increasing if $\lambda \leqslant 1$ and R is odd if ϕ is even.*

(iii) *Suppose that $\mathcal{H} = \mathbb{R}$ and that R is increasing. Then there exists $\phi \in \Gamma_0(\mathbb{R})$ such that $R = \operatorname{prox}_\phi$.*

Next, we illustrate the pervasiveness of nonexpansive averaged activation operators in practice, starting with activation operators on the real line.

Example 3.2 Proposition 3.1(ii) states that activation functions on the real line can be expressed in the generic form

$$R = \operatorname{Id} + \lambda(\operatorname{prox}_\phi - \operatorname{Id}), \quad \text{where} \quad \phi \in \Gamma_0(\mathbb{R}) \quad \text{and} \quad \lambda \in [0,2]. \tag{3.4}$$

Here are a few explicit instantiations of this proximal representation.

(i) If $\lambda = 1$, we obtain the class of proximal activation functions discussed in [24] and which was seen there to include standard instances such as the unimodal sigmoid activation function [24, Example 2.13], the saturated linear activation function [24, Example 2.5], the ReLU activation function [24, Example 2.6], the inverse square root unit activation function [24, Example 2.9], the hyperbolic tangent activation function [24, Example 2.12], and the Elliot activation function [24, Example 2.15]. Additional examples in this category are the following. Given $\beta \in \left]0, +\infty\right[$, the capped ReLU activation function [36] is

$$(\forall x \in \mathbb{R}) \quad R(x) = \operatorname{prox}_{\iota_{[0,\beta]}}(x) = \min\{\max\{x, 0\}, \beta\}, \tag{3.5}$$

and, for $\beta \leqslant 1$, the exponential linear unit (ELU) function [20] is

$$(\forall x \in \mathbb{R}) \quad R(x) = \begin{cases} x, & \text{if } x \geqslant 0; \\ \beta\big(\exp(x) - 1\big), & \text{if } x < 0. \end{cases} \tag{3.6}$$

It follows from [8, Cor. 24.5, Prop. 24.32, and Exa. 13.2(v)] that $R = \operatorname{prox}_\phi$, where

$$(\forall x \in \mathbb{R}) \quad \phi(x) = \begin{cases} 0, & \text{if } x \geqslant 0; \\ (x+\beta)\ln\left(\dfrac{x+\beta}{\beta}\right) - x - \dfrac{x^2}{2}, & \text{if } -\beta < x < 0; \\ \beta - \dfrac{\beta^2}{2}, & \text{if } x = -\beta; \\ +\infty, & \text{if } x < -\beta. \end{cases} \tag{3.7}$$

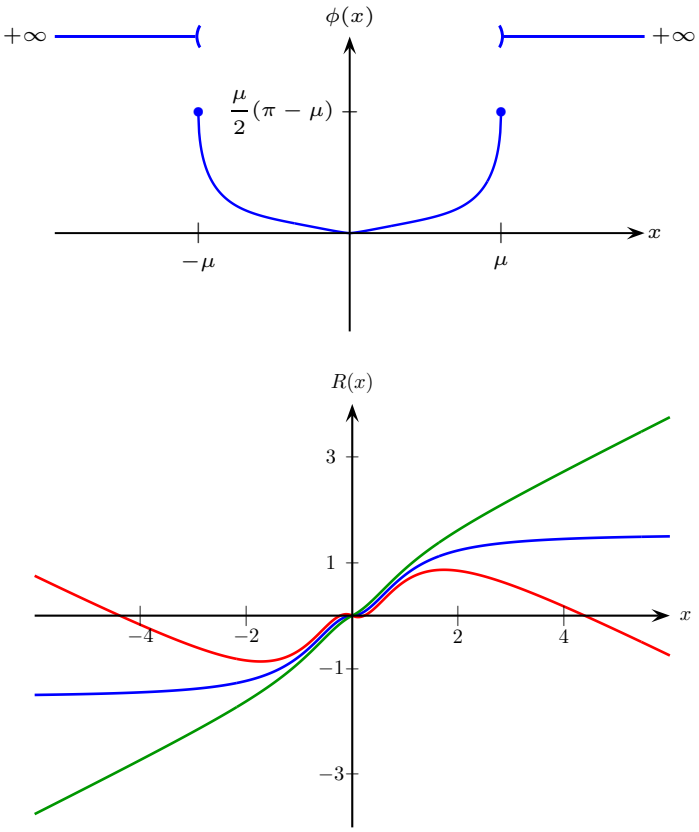

Figure 2: Averaged activation functions: Illustration of Example 3.2(ii). Top: The function ϕ of (3.10). Bottom: In blue, the activation operator R of (3.8) is the proximity operator of ϕ, which corresponds to $\lambda = 1$ in (3.4). The green curve corresponds to the case when $\lambda = 0.5$ in (3.4), and the red one to the case when $\lambda = 1.5$. As stated in Proposition 3.1(i), relaxation parameters $\lambda \in [0, 1]$ yield increasing activation functions. Non-monotonic averaged activation functions in (3.4) must be generated with relaxation parameters $\lambda \in]1, 2]$. As seen in Proposition 3.1(ii), since ϕ is even, R is odd.

The softplus activation function [29] $R\colon x \mapsto \ln((1+e^x)/2)$ is also a proximity operator since it is nonexpansive and increasing (see Proposition 3.1(iii)).

(ii) The Geman–McClure function [28]

$$(\forall x \in \mathbb{R}) \quad R(x) = \frac{\mu \operatorname{sign}(x) x^2}{1+x^2}, \quad \text{where} \quad \mu = \frac{8}{3\sqrt{3}}, \tag{3.8}$$

will be employed in Example 3.3. Set $\psi = |\cdot| - \arctan|\cdot| \in \Gamma_0(\mathbb{R})$. Then R is nonexpansive and $R = \mu\psi'$. The conjugate of $\mu\psi$ is 1-strongly convex and given by $\mu\psi^*(\cdot/\mu)$, where

$$(\forall x \in \mathbb{R}) \quad \psi^*(x) = \begin{cases} \arctan\sqrt{\dfrac{|x|}{1-|x|}} - \sqrt{|x|(1-|x|)}, & \text{if } |x| < 1; \\ \dfrac{\pi}{2}, & \text{if } |x| = 1; \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.9}$$

It follows from [8, Cor. 24.5] that $R = \operatorname{prox}_\phi$ with (see Fig. 2)

$$\phi = \mu\psi^*\left(\frac{\cdot}{\mu}\right) - \frac{|\cdot|^2}{2} \colon x \mapsto \begin{cases} \mu \arctan\sqrt{\dfrac{|x|}{\mu - |x|}} - \sqrt{|x|(\mu - |x|)} - \dfrac{x^2}{2}, & \text{if } |x| < \mu; \\ \dfrac{\mu(\pi - \mu)}{2}, & \text{if } |x| = \mu; \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.10}$$

(iii) Take $\phi = \iota_{[0,+\infty[}$. Then we obtain the leaky ReLU activation function [38] for $0 < \lambda < 1$, the ReLU activation function for $\lambda = 1$, and the absolute value activation function [17] for $\lambda = 2$.

(iv) The use of nonmonotonic activation functions has been advocated in various papers. They turn out to be α-averaged (alternatively, in view of Proposition 3.1(ii), they are of the form (3.4) with $\lambda \in]1, 2]$). To compute the averagedness constant of a nonexpansive operator $R \colon \mathbb{R} \to \mathbb{R}$, one can proceed as follows. According to (1.2), we must find the smallest $\alpha \in]0, 1]$ such that $Q = \operatorname{Id} + \alpha^{-1}(R - \operatorname{Id})$ remains nonexpansive. This means that the supremum of the modulus of the one-sided derivatives (the derivatives if they exist) over $\mathbb{R}$ should be one. Thus, we obtain $\alpha = 1$ for the sine activation function $R = \sin$ [42], as well as for the absolute value function $R = |\cdot|$ [17] and the mirrored ReLU activation function [58]

$$(\forall x \in \mathbb{R}) \quad R(x) = \operatorname{proj}_{[0,1]}|x| = \begin{cases} |x|, & \text{if } |x| < 1; \\ 1, & \text{otherwise,} \end{cases} \tag{3.11}$$

$\alpha \approx 0.546$ for the swish activation function [45]

$$(\forall x \in \mathbb{R}) \quad R(x) = \frac{10x}{11(1 + \exp(-x))}, \tag{3.12}$$

$\alpha \approx 0.536$ for the exponential linear squashing (ELiSH) function [7]

$$(\forall x \in \mathbb{R}) \quad R(x) = \frac{10}{11} \times \begin{cases} \dfrac{x}{1 + \exp(-x)}, & \text{if } x \geqslant 0; \\ \dfrac{\exp(x) - 1}{1 + \exp(-x)}, & \text{if } x < 0, \end{cases} \tag{3.13}$$

and $\alpha = (1 + \sqrt{2/e})/2$ for the Gaussian activation function $R \colon x \mapsto \exp(-x^2)$ [40].

Next, is a technique for lifting a proximal activation operator from $\mathbb{R}$ to a Hilbert space $\mathcal{H}$.

Example 3.3 Let $\mathcal{H}$ be a real Hilbert space, let $\lambda \in [0, 2]$, let C be a nonempty closed convex subset of $\mathcal{H}$, let $\phi \in \Gamma_0(\mathbb{R})$ be an even function such that ϕ^* is differentiable on $\mathbb{R} \smallsetminus \{0\}$ with 0 as its unique minimizer. Set

$$(\forall x \in \mathcal{H}) \quad Rx = \begin{cases} (1 - \lambda)x + \dfrac{\lambda \operatorname{prox}_\phi d_C(x)}{d_C(x)}(x - \operatorname{proj}_C x), & \text{if } x \notin C; \\ (1 - \lambda)x, & \text{if } x \in C. \end{cases} \tag{3.14}$$

Then R is $\lambda/2$-averaged. In particular, set $\lambda = 1$, $C = \{0\}$, $\mu = 8/(3\sqrt{3})$ and define ϕ as in (3.10). Then we infer that the squashing function

$$R \colon x \mapsto \frac{\mu\|x\|}{1 + \|x\|^2}x \tag{3.15}$$

used in capsule networks [48] is a proximal activation operator.

Another construction that builds on activation functions on the real line is the following, which is reminiscent of the original multilayer perceptrons [46].

Example 3.4 Let $\mathcal{H}$ be a separable real Hilbert space, let $\varnothing \neq \mathbb{K} \subset \mathbb{N}$, let $(e_k)_{k\in\mathbb{K}}$ be an orthonormal basis of $\mathcal{H}$, and let $\alpha \in [0,1]$. For every $k \in \mathbb{K}$, let $\varrho_k \colon \mathbb{R} \to \mathbb{R}$ be α-averaged and such that $\varrho_k(0) = 0$. Define $R \colon \mathcal{H} \to \mathcal{H} \colon x \mapsto \sum_{k\in\mathbb{K}} \varrho_k(\langle x \mid e_k\rangle)e_k$. Then R is α-averaged.

Example 3.5 Let N be a strictly positive integer, let $\omega \in [0,1]$, and let C be a nonempty closed convex subset of $\mathbb{R}^N$. Set

$$R \colon \mathbb{R}^N \to \mathbb{R}^N \colon (\xi_k)_{1\leqslant k\leqslant N} \mapsto \omega\big(\xi_k^{\uparrow}\big)_{1\leqslant k\leqslant N} + (1-\omega)\mathrm{proj}_C(\xi_k)_{1\leqslant k\leqslant N}, \tag{3.16}$$

where $(\xi_k^{\uparrow})_{1\leqslant k\leqslant N}$ denotes the vector obtained by sorting the components of $(\xi_k)_{1\leqslant k\leqslant N}$ in ascending order. Then R is $(1+\omega)/2$-averaged.

Remark 3.6 Set $C = \big\{(\xi_k)_{1\leqslant k\leqslant N} \in \mathbb{R}^N \;\big|\; \xi_1 = \cdots = \xi_N\big\}$ in Example 3.5. Then

$$R \colon \mathbb{R}^N \to \mathbb{R}^N \colon (\xi_k)_{1\leqslant k\leqslant N} \mapsto \left(\omega\xi_k^{\uparrow} + \frac{1-\omega}{N}\sum_{j=1}^{N}\xi_j\right)_{1\leqslant k\leqslant N}. \tag{3.17}$$

Now set $W \colon \mathbb{R}^N \to \mathbb{R} \colon (\xi_k)_{1\leqslant k\leqslant N} \mapsto \xi_N$. Then $W \circ R$ corresponds to the max-average pooling performed on a block of size N [37]. When $\omega = 0$, the standard average-pooling operation is obtained, which is associated with the activation operator proj_C. When $\omega = 1$, we recover the standard max-pooling operation [14], which is the main building block of maxout layers [31]. The max-pooling operator is nonexpansive.

Example 3.7 Let $2 \leqslant N \in \mathbb{N}$, let $\{\tau_k\}_{1\leqslant k\leqslant N-1} \subset \,]-1,1[$, and let $\theta \in \mathbb{R}$. Set

$$R \colon \mathbb{R}^{N-1} \to \mathbb{R}^{N-1} \colon (\xi_k)_{1\leqslant k\leqslant N-1} \mapsto US\Big([\tau_1\xi_1, \ldots, \tau_{N-1}\xi_{N-1}, \theta]^{\top}\Big), \tag{3.18}$$

where $U \in \mathbb{R}^{(N-1)\times N}$ is the matrix obtained by retaining the first $(N-1)$ rows of the identity matrix of size $N \times N$, and $S \colon \mathbb{R}^N \to \mathbb{R}^N \colon (\xi_k)_{1\leqslant k\leqslant N} \mapsto (\xi_k^{\uparrow})_{1\leqslant k\leqslant N}$. Then R is $(1+\max\{|\tau_1|, \ldots, |\tau_{N-1}|\})/2$-averaged.

Remark 3.8 Let $N \geqslant 3$ be an odd integer, let $(\tau_k)_{1\leqslant k\leqslant N-1} \in \,]-1,1[^{N-1}$, let $\theta \in \mathbb{R}$, let R be the activation operator defined in Example 3.7, and set $W \colon \mathbb{R}^{N-1} \to \mathbb{R} \colon (\xi_k)_{1\leqslant k\leqslant N-1} \mapsto \xi_{\frac{N+1}{2}}$. Then, for every $x = (\xi_k)_{1\leqslant k\leqslant N-1} \in \mathbb{R}^{N-1}$, $(W \circ R)x = \mathrm{median}\{\tau_1\xi_1, \ldots, \tau_{N-1}\xi_{N-1}, \theta\}$. This corresponds to the median neuron model introduced in [2].

Remark 3.9 Multi-component averaged activation operators can be derived from the above examples. Indeed, let $(\mathcal{H}_i)_{1\leqslant i\leqslant M}$ be real Hilbert spaces and let $\mathcal{H} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_M$ be their Hilbert direct sum. For every $i \in \{1, \ldots, M\}$, let $\alpha_i \in [0,1]$ and let $R_i \colon \mathcal{H}_i \to \mathcal{H}_i$ be α_i-averaged. Then $R \colon \mathcal{H} \to \mathcal{H} \colon (x_i)_{1\leqslant i\leqslant M} \mapsto (R_i x_i)_{1\leqslant i\leqslant M}$ is α-averaged with $\alpha = \max_{1\leqslant i\leqslant M}\alpha_i$.

4 Lipschitz constants for layered networks

The objective of this section is to derive Lipschitz constants for networks conforming to Model 1.1. Note that, if $m = 1$, a Lipschitz constant of T is clearly $\theta_1 = \|W_1\|$ since R_1 is nonexpansive. We shall therefore focus henceforth on the case $m \geqslant 2$. Throughout, the following notation is employed.

Notation 4.1 Let $2 \leqslant m \in \mathbb{N}$ and $k \in \{1, \ldots, m-1\}$. Then

$$\mathbb{J}_{m,k} = \begin{cases} \big\{(j_1, \ldots, j_k) \in \mathbb{N}^k \;\big|\; 1 \leqslant j_1 < \cdots < j_k \leqslant m-1\big\}, & \text{if } k > 1; \\ \{1, \ldots, m-1\}, & \text{if } k = 1 \end{cases} \tag{4.1}$$

and, for every $(j_1, \ldots, j_k) \in \mathbb{J}_{m,k}$,

$$\sigma_{m;\{j_1,\ldots,j_k\}} = \|W_m \circ \cdots \circ W_{j_k+1}\| \, \|W_{j_k} \circ \cdots \circ W_{j_{k-1}+1}\| \cdots \|W_{j_1} \circ \cdots \circ W_1\|. \tag{4.2}$$

Theorem 4.2 *Consider the setting of Model 1.1 with $m \geqslant 2$. Set*

$$(\forall\, \mathbb{J} \subset \{1, \ldots, m-1\}) \quad \beta_{m;\mathbb{J}} = \bigg(\prod_{j \in \mathbb{J}} \alpha_j\bigg) \prod_{j \in \{1,\ldots,m-1\} \smallsetminus \mathbb{J}} (1 - \alpha_j) \tag{4.3}$$

and

$$\theta_m = \beta_{m;\varnothing} \|W_m \circ \cdots \circ W_1\| + \sum_{k=1}^{m-1} \sum_{(j_1,\ldots,j_k) \in \mathbb{J}_{m,k}} \beta_{m;\{j_1,\ldots,j_k\}} \sigma_{m;\{j_1,\ldots,j_k\}}. \tag{4.4}$$

Then θ_m is a Lipschitz constant of T.

The following proposition features some important special cases.

Proposition 4.3 *Consider the setting of Model 1.1 with $m \geqslant 2$, and let θ_m be defined as in (4.4). Then the following hold:*

(i) $\|W_m \circ \cdots \circ W_1\| \leqslant \theta_m \leqslant \prod_{i=1}^m \|W_i\|$.

(ii) *Suppose that, for every $i \in \{1, \ldots, m-1\}$, $R_i = \mathrm{Id}$. Then $\theta_m = \|W_m \circ \cdots \circ W_1\|$.*

(iii) *Suppose that, for every $i \in \{1, \ldots, m-1\}$, R_i is purely nonexpansive in the sense that $\alpha_i = 1$ is its smallest averaging constant. Then $\theta_m = \prod_{i=1}^m \|W_i\|$.*

(iv) *Suppose that, for every $i \in \{1, \ldots, m-1\}$, R_i is firmly nonexpansive. Then*

$$\theta_m = \frac{1}{2^{m-1}} \bigg(\|W_m \circ \cdots \circ W_1\| + \sum_{k=1}^{m-1} \sum_{(j_1,\ldots,j_k) \in \mathbb{J}_{m,k}} \sigma_{m;\{j_1,\ldots,j_k\}}\bigg). \tag{4.5}$$

(v) *Set $\alpha_0 = \theta_0 = 1$. Then*

$$\theta_m = \sum_{i=0}^{m-1} \alpha_i \theta_i \bigg(\prod_{q=i+1}^{m-1} (1 - \alpha_q)\bigg) \|W_m \circ \cdots \circ W_{i+1}\|. \tag{4.6}$$

Remark 4.4 Proposition 4.3(i)–4.3(iii) show that the tightest bound in terms of stability corresponds to a linear network, while the loosest corresponds to a network with nonlinearities having no stronger property than nonexpansiveness.

We close this section by observing that the Lipschitz constant exhibited in Theorem 4.2 is a componentwise increasing function of the averagedness constants of the activation operators.

Proposition 4.5 *Consider the setting of Model 1.1 with $m \geqslant 2$. Make the Lipschitz constant θ_m in Theorem 4.2 a function of $(\alpha_1, \ldots, \alpha_{m-1}) \in [0,1]^{m-1}$. Let $(\alpha_i)_{1 \leqslant i \leqslant m-1} \in [0,1]^{m-1}$ and $(\alpha'_i)_{1 \leqslant i \leqslant m-1} \in [0,1]^{m-1}$ be such that $(\forall i \in \{1, \ldots, m-1\})\ \alpha_i \leqslant \alpha'_i$. Then $\theta_m(\alpha_1, \ldots, \alpha_{m-1}) \leqslant \theta_m(\alpha'_1, \ldots, \alpha'_{m-1})$.*

Remark 4.6 Proposition 4.5 suggests that, in terms of stability, it is better to use proximal activation operators, such as those listed in Example 3.2(i)–(ii), than α-averaged activation operators for which $\alpha > 1/2$, such as those mentioned in Example 3.2(iv).

5 Networks using separable activation operators

We show that sharper Lipschitz constants can be derived in the case of networks featuring the type of separable structure described in Example 3.4. Note that this class of networks is the most commonly used, standard convnets being special cases. The following notation will be used.

Notation 5.1 Let $\mathcal{H}$ be a separable real Hilbert space, let $\varnothing \neq \mathbb{K} \subset \mathbb{N}$, let $E = (e_k)_{k\in\mathbb{K}}$ be an orthonormal basis of $\mathcal{H}$, and let I be a nonempty bounded subset of $\mathbb{R}$. Then

$$\mathscr{D}_I(E) = \left\{ \Lambda\colon \mathcal{H} \to \mathcal{H}\colon x \mapsto \sum_{k\in\mathbb{K}} \lambda_k \langle x \mid e_k\rangle e_k \;\middle|\; \{\lambda_k\}_{k\in\mathbb{K}} \subset I \right\}. \tag{5.1}$$

5.1 General results

Theorem 5.2 *Consider the setting of Model 1.1 with $m \geqslant 2$. For every $i \in \{1,\ldots,m-1\}$, suppose that $\mathcal{H}_i$ is separable, let $\varnothing \neq \mathbb{K}_i \subset \mathbb{N}$, let $E_i = (e_{i,k})_{k\in\mathbb{K}_i}$ be an orthonormal basis of $\mathcal{H}_i$, and, for every $k \in \mathbb{K}_i$, let $\varrho_{i,k}\colon \mathbb{R}\to\mathbb{R}$ be α_i-averaged and such that $\varrho_{i,k}(0) = 0$. Assume that*

$$(\forall i \in \{1,\ldots,m-1\}) \quad R_i\colon \mathcal{H}_i \to \mathcal{H}_i\colon x \mapsto \sum_{k\in\mathbb{K}_i} \varrho_{i,k}(\langle x \mid e_{i,k}\rangle) e_{i,k} \tag{5.2}$$

and define

$$\vartheta_m = \sup_{\substack{\Lambda_1\in\mathscr{D}_{\{1-2\alpha_1,1\}}(E_1)\\ \vdots \\ \Lambda_{m-1}\in\mathscr{D}_{\{1-2\alpha_{m-1},1\}}(E_{m-1})}} \|W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\|. \tag{5.3}$$

Then the following hold:

(i) *ϑ_m is a Lipschitz constant of the operator T of (1.3).*

(ii) *Define θ_m as in (4.4). Then $\|W_m\circ\cdots\circ W_1\| \leqslant \vartheta_m \leqslant \theta_m$.*

Remark 5.3 An expression similar to (5.3) was proposed empirically in [49] for a multilayer perceptron operating on finite-dimensional spaces under the additional assumption that the activation operators are continuously differentiable and firmly nonexpansive.

Remark 5.4 In Theorem 5.2, make the additional assumption that, for some $i \in \{1,\ldots,m-1\}$, the functions $(\varrho_{i,k})_{k\in\mathbb{K}_i}$ are increasing. Then it follows from Proposition 3.1(iii) that there exist functions $(\phi_{i,k})_{k\in\mathbb{K}_i}$ in $\Gamma_0(\mathbb{R})$ such that $(\forall k \in \mathbb{K}_i)\ \varrho_{i,k} = \operatorname{prox}_{\phi_{i,k}}$. In addition, for every $k \in \mathbb{K}_i$, since $\varrho_{i,k}(0) = 0$ and since the set of minimizers of $\phi_{i,k}$ coincides with the set of fixed points of $\operatorname{prox}_{\phi_{i,k}}$ [8, Proposition 12.29], we deduce that $\phi_{i,k}$ is minimized at 0. Furthermore, $\alpha_i = 1/2$ and $R_i = \operatorname{prox}_{\varphi_i}$, where $(\forall x \in \mathcal{H}_i)\ \varphi_i(x) = \sum_{k\in\mathbb{K}_i} \phi_{i,k}(\langle x \mid e_{i,k}\rangle)$. Such a construction is used in [23, 25].

As in Proposition 4.5, the Lipschitz constant exhibited in Theorem 5.2 turns out to be a componentwise increasing function of the averagedness constants of the activation operators.

Proposition 5.5 *Consider the setting of Model 1.1 with $m \geqslant 2$. For every $i \in \{1, \dots, m-1\}$, suppose that $\mathcal{H}_i$ is separable, let $\varnothing \neq \mathbb{K}_i \subset \mathbb{N}$, and let $E_i = (e_{i,k})_{k \in \mathbb{K}_i}$ be an orthonormal basis of $\mathcal{H}_i$. Define $\vartheta_m \colon [0,1]^{m-1} \to [0, +\infty[$ by*

$$\vartheta_m \colon (\alpha_1, \dots, \alpha_{m-1}) \mapsto \sup_{\substack{\Lambda_1 \in \mathscr{D}_{\{1-2\alpha_1,1\}}(E_1) \\ \vdots \\ \Lambda_{m-1} \in \mathscr{D}_{\{1-2\alpha_{m-1},1\}}(E_{m-1})}} \|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|. \tag{5.4}$$

Let $(\alpha_i)_{1 \leqslant i \leqslant m-1} \in [0,1]^{m-1}$ and $(\alpha'_i)_{1 \leqslant i \leqslant m-1} \in [0,1]^{m-1}$ be such that $(\forall i \in \{1, \dots, m-1\})$ $\alpha_i \leqslant \alpha'_i$. Then $\vartheta_m(\alpha_1, \dots, \alpha_{m-1}) \leqslant \vartheta_m(\alpha'_1, \dots, \alpha'_{m-1})$.

5.2 Extension to non-Hilbertian norms

In certain applications, Hilbertian norms may not be the most relevant measures to quantify errors. We now state a variant of Theorem 5.2 which holds for alternative norms. It involves embeddings of Hilbert spaces; standard examples can be found in [57]. Let us also point out that these embedding conditions are automatically satisfied if the spaces are finite-dimensional.

Proposition 5.6 *Consider the setting of Model 1.1 with $m \geqslant 2$. For every $i \in \{1, \dots, m\}$, suppose that $\mathcal{H}_i$ is separable, let $\varnothing \neq \mathbb{K}_i \subset \mathbb{N}$, let $E_i = (e_{i,k})_{k \in \mathbb{K}_i}$ be an orthonormal basis of $\mathcal{H}_i$, and, for every $k \in \mathbb{K}_i$, let $\varrho_{i,k} \colon \mathbb{R} \to \mathbb{R}$ be α_i-averaged and such that $\varrho_{i,k}(0) = 0$. Let $\mathcal{G}_0$ be the normed space obtained by equipping the vector space underlying $\mathcal{H}_0$ with a norm for which $\mathcal{G}_0$ is continuously embedded in $\mathcal{H}_0$, and let $\mathcal{G}_m$ be the normed space obtained by equipping the vector space underlying $\mathcal{H}_m$ with a norm for which $\mathcal{H}_m$ is continuously embedded in $\mathcal{G}_m$. Assume that*

$$(\forall i \in \{1, \dots, m\}) \quad R_i \colon \mathcal{H}_i \to \mathcal{H}_i \colon x \mapsto \sum_{k \in \mathbb{K}_i} \varrho_{i,k}\big(\langle x \mid e_{i,k} \rangle\big) e_{i,k}. \tag{5.5}$$

Then

$$\vartheta_m = \sup_{\substack{\Lambda_1 \in \mathscr{D}_{\{1-2\alpha_1,1\}}(E_1) \\ \vdots \\ \Lambda_m \in \mathscr{D}_{\{1-2\alpha_m,1\}}(E_m)}} \|\Lambda_m \circ W_m \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0, \mathcal{G}_m} \tag{5.6}$$

is a Lipschitz constant of $T \colon \mathcal{G}_0 \to \mathcal{G}_m$.

Corollary 5.7 *Consider the setting of Model 1.1 with $m \geqslant 2$. Define $\mathcal{G}_0$ and $(R_i)_{1 \leqslant i \leqslant m}$ as in Proposition 5.6, let $p \in [1, +\infty]$, and let $\{\omega_k\}_{k \in \mathbb{K}_m} \subset]0, +\infty[$ be such that one of the following holds:*

(i) $p \in [1, 2[$ *and* $\sum_{k \in \mathbb{K}_m} \omega_k^{2/(2-p)} < +\infty$.

(ii) $p \in [2, +\infty]$ *and* $\sup_{k \in \mathbb{K}_m} \omega_k < +\infty$.

Let $\mathcal{G}_m$ be the normed space obtained by equipping the vector space underlying $\mathcal{H}_m$ with the norm

$$\big(\forall x \in \mathcal{H}_m\big) \quad \|x\|_{\mathcal{G}_m} = \begin{cases} \left| \sum_{k \in \mathbb{K}_m} \omega_k |\langle x \mid e_{m,k} \rangle|^p \right|^{1/p}, & \text{if } p < +\infty; \\ \sup_{k \in \mathbb{K}_m} \omega_k |\langle x \mid e_{m,k} \rangle|, & \text{if } p = +\infty. \end{cases} \tag{5.7}$$

Then a Lipschitz constant of $T\colon \mathcal{G}_0 \to \mathcal{G}_m$ *is*

$$\vartheta_m = \sup_{\substack{\Lambda_1 \in \mathscr{D}_{\{1-2\alpha_1,1\}}(E_1) \\ \vdots \\ \Lambda_{m-1} \in \mathscr{D}_{\{1-2\alpha_{m-1},1\}}(E_{m-1})}} \|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}. \tag{5.8}$$

5.3 Networks with positive weights

Under certain positivity assumptions, the constant ϑ_m of (5.3) and (5.8) can be simplified.

Assumption 5.8 Consider the setting of Model 1.1 with $m \geqslant 2$. For every $i \in \{0,\ldots,m\}$, suppose that $\mathcal{H}_i$ is separable, let $\varnothing \neq \mathbb{K}_i \subset \mathbb{N}$, and let $E_i = (e_{i,k})_{k\in\mathbb{K}_i}$ be an orthonormal basis of $\mathcal{H}_i$. For every $(k_0,\ldots,k_m) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_m$, set

$$\mu_{k_0,\ldots,k_m} = \langle W_1 e_{0,k_0} \mid e_{1,k_1}\rangle \cdots \big\langle W_m e_{m-1,k_{m-1}} \mid e_{m,k_m}\big\rangle. \tag{5.9}$$

We suppose that

$$(\forall (k_0,\ldots,k_m) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_m)(\forall (l_0,\ldots,l_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1}) \\ \mu_{k_0,\ldots,k_{m-1},k_m}\, \mu_{l_0,\ldots,l_{m-1},k_m} \geqslant 0. \tag{5.10}$$

Example 5.9 Consider the particular case of Model 1.1 in which, for every $i \in \{0,\ldots,m\}$, $N_i \in \mathbb{N}\smallsetminus\{0\}$, $\mathcal{H}_i = \mathbb{R}^{N_i}$, E_i is the canonical basis of $\mathbb{R}^{N_i}$ and, for every $k \in \{1,\ldots,N_i\}$, $\chi_{i,k} \in \{-1,1\}$ with the additional condition that, for every $l \in \{1,\ldots,N_0\}$, $\chi_{0,k} = \chi_{0,l}$. Further, for every $i \in \{1,\ldots,m\}$, the matrix $W_i = [w_{i,k,l}]_{1\leqslant k\leqslant N_i, 1\leqslant l\leqslant N_{i-1}} \in \mathbb{R}^{N_i\times N_{i-1}}$ satisfies

$$(\forall k \in \{1,\ldots,N_i\})(\forall l \in \{1,\ldots,N_{i-1}\}) \quad w_{i,k,l} = \chi_{i,k}\chi_{i-1,l}|w_{i,k,l}|. \tag{5.11}$$

Then Assumption 5.8 holds. This is true in particular if, for every $i \in \{1,\ldots,m\}$, $\{w_{i,k,l}\}_{1\leqslant k\leqslant N_i, 1\leqslant l\leqslant N_{i-1}} \subset [0,+\infty[$, which corresponds to positively weighted networks. See [19] for the design of such networks.

In the following result, a Lipschitz constant of the network (1.3) coincides with that of the linear network $W_m \circ \cdots \circ W_1$ for standard choices of norms.

Proposition 5.10 *Suppose that the assumptions of Corollary 5.7 are satisfied, that*

$$\big(\forall (\xi_k)_{k\in\mathbb{K}_0} \in \ell^2(\mathbb{K}_0)\big) \quad \bigg\| \sum_{k\in\mathbb{K}_0} \xi_k e_{0,k} \bigg\|_{\mathcal{G}_0} = \bigg\| \sum_{k\in\mathbb{K}_0} |\xi_k| e_{0,k} \bigg\|_{\mathcal{G}_0}, \tag{5.12}$$

and that Assumption 5.8 holds. Then the Lipschitz constant ϑ_m *of* $T\colon \mathcal{G}_0 \to \mathcal{G}_m$ *in* (5.8) *reduces to* $\vartheta_m = \|W_m \circ \cdots \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}$.

We show below that the Lipschitz constant of a positively weighted network associated with weight operators $(W_i)_{1\leqslant i\leqslant m}$ and nonseparable activation operators is not necessarily $\|W_m \circ \cdots \circ W_1\|$.

Example 5.11 Consider the toy version of Model 1.1 in which $m = 2$, $\mathcal{H}_0 = \mathcal{H}_1 = \mathcal{H}_2 = \mathbb{R}^2$. Set $\varphi\colon x = (\xi_1,\xi_2) \mapsto \phi(\xi_1) + \phi(\xi_2)$, where

$$\phi\colon \mathbb{R} \to \left]-\infty,+\infty\right] : \xi \mapsto \begin{cases} \dfrac{(1+\xi)\ln(1+\xi) + (1-\xi)\ln(1-\xi) - \xi^2}{2} & \text{if } |\xi| < 1; \\ \ln(2) - 1/2 & \text{if } |\xi| = 1; \\ +\infty, & \text{if } |\xi| > 1. \end{cases} \tag{5.13}$$

Let $\xi \in]-1,1[= \operatorname{dom}\phi' = \operatorname{dom}(\operatorname{Id}+\phi') = \operatorname{ran}\operatorname{prox}_{\phi}$. Then $\xi + \phi'(\xi) = \operatorname{arctanh}(\xi)$ and therefore $\varrho = (\operatorname{Id}+\phi')^{-1} = \tanh$. Consequently, we derive from [24, Example 2.13] that $(\forall x = (\xi_1, \xi_2) \in \mathbb{R}^2)$ $\operatorname{prox}_{\varphi} x = \big(\tanh(\xi_1), \tanh(\xi_2)\big)$. Now set

$$b_1 = b_2 = 0, \quad U = \frac{1}{2}\begin{bmatrix} \sqrt{3} & 1 \\ 1 & -\sqrt{3} \end{bmatrix}, \quad W_1 = \begin{bmatrix} 1 & 3 \\ 3 & 3 \end{bmatrix}, \quad W_2 = \begin{bmatrix} 10 & 2 \\ 7 & 4 \end{bmatrix}, \tag{5.14}$$

$R_1 = \operatorname{prox}_{\varphi\circ U} = U \circ \operatorname{prox}_{\varphi} \circ U$ [25, Lemma 2.8], and $R_2 = \operatorname{Id}$. Then $\|W_2W_1\| \approx 54.72$. If the input $x = (-3.4, 2)$ is perturbed by $z = 10^{-4} \times (1, \sqrt{3})$, we get $\|T(x+z) - Tx\|/\|z\| \approx 58.18$, which shows that, although W_1 and W_2 have strictly positive entries, the Lipschitz constant is larger than $\|W_2W_1\|$. Note that, in this scenario, the constant of (4.4) is

$$\theta_2 = (\|W_2W_1\| + \|W_2\|\|W_1\|)/2 \approx 60.50. \tag{5.15}$$

A sharper Lipschitz constant can be obtained by noticing that this network is equivalent to a network in which W_1, W_2, and R_1 are replaced by $W_1' = UW_1$, $W_2' = W_2U$, and $R_1' = \operatorname{prox}_{\varphi}$. Since R_1' is separable, the constant of (5.4) is $\vartheta_2 \approx 59.54$. In contrast, the naive bound of (1.4) is about 66.29.

For separable activators in finite-dimensional spaces, we have the following result, which does not require Assumption 5.8.

Proposition 5.12 *Consider the setting of Model 1.1 with $m \geqslant 2$. Suppose that the assumptions of Corollary 5.7 hold and that $\|\cdot\|_{\mathcal{G}_0}$ satisfies (5.12). In addition, assume that, for every $i \in \{0,\ldots,m\}$, $\mathcal{H}_i = \mathbb{R}^{N_i}$ and E_i is the canonical basis of $\mathbb{R}^{N_i}$. For every $i \in \{1,\ldots,m\}$, let A_i denote the matrix obtained by taking the absolute values of the entries of the matrix W_i. Then the Lipschitz constant ϑ_m of $T\colon \mathcal{G}_0 \to \mathcal{G}_m$ in (5.8) satisfies $\vartheta_m \leqslant \|A_m \cdots A_1\|_{\mathcal{G}_0,\mathcal{G}_m}$.*

6 Conclusion

Using advanced tools from nonlinear analysis, we have derived sharp Lipschitz constants for layered network structures involving compositions of nonexpansive averaged operators and affine operators. This framework has been shown to model feed-forward neural networks having a chain graph structure. Extending these results to networks having a more general dyadic acyclic graph (DAG) structure would be of interest. Among the many avenues of future research that this work suggests, it would be interesting to exploit it to devise training strategies to achieve better robustness. The proposed nonexpansive operator machinery could also be used to design network architectures with smaller Lipschitz constants. Finally, computing local Lipschitz constants could be of interest in practice and constitutes an important topic of future research.

A Technical lemmas

Lemma A.1 [23, Proposition 2.4] *Let R be a function defined from $\mathbb{R}$ to $\mathbb{R}$. Then R is the proximity operator of a function in $\Gamma_0(\mathbb{R})$ if and only if it is nonexpansive and increasing.*

Lemma A.2 *Let $q \in \mathbb{N} \smallsetminus \{0\}$ and, for every $i \in \{1,\ldots,q\}$, let S_i be a nonempty subset of a real vector space $\mathcal{X}_i$. Let $\psi\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_q \to \mathbb{R}$ be a function which is convex with respect to each of its q coordinates. Set $\boldsymbol{S} = S_1 \times \cdots \times S_q$ and let $\operatorname{conv}\boldsymbol{S}$ be its convex envelope. Then $\sup\psi(\boldsymbol{S}) = \sup\psi(\operatorname{conv}\boldsymbol{S})$.*

Proof. Set $\mu = \sup \psi(\boldsymbol{S})$. Clearly, $\mu \leqslant \sup \psi(\operatorname{conv} \boldsymbol{S})$. Now take $\boldsymbol{x} \in \operatorname{conv} \boldsymbol{S}$. Then $\boldsymbol{x} = \sum_{j\in I} \alpha_j \boldsymbol{x}_j$, where $(\alpha_j)_{j\in I}$ is a finite family in $]0,1]$ such that $\sum_{j\in I} \alpha_j = 1$ and, for every $j \in I$, $\boldsymbol{x}_j = (x_{j,i})_{1\leqslant i\leqslant q}$, with $(\forall i \in \{1,\ldots,q\})\ x_{j,i} \in S_i$. Note that $(\forall (j_1,\ldots,j_q) \in I^q)\ (x_{j_1,1},\ldots,x_{j_q,q}) \in \boldsymbol{S}$. Therefore,

$$\begin{aligned}
\psi(\boldsymbol{x}) &= \psi\bigg(\sum_{j_1\in I} \alpha_{j_1} x_{j_1,1}, \ldots, \sum_{j_q\in I} \alpha_{j_q} x_{j_q,q}\bigg) \\
&\leqslant \sum_{j_1\in I} \alpha_{j_1} \psi\bigg(x_{j_1,1}, \sum_{j_2\in I} \alpha_{j_2} x_{j_2,2}, \ldots, \sum_{j_q\in I} \alpha_{j_q} x_{j_q,q}\bigg) \\
&\;\;\vdots \\
&\leqslant \sum_{j_1\in I,\ldots,j_q\in I} \bigg(\prod_{i=1}^{q} \alpha_{j_i}\bigg) \psi\big(x_{j_1,1},\ldots,x_{j_q,q}\big) \leqslant \mu.
\end{aligned} \tag{A.1}$$

Hence, $\sup \psi(\operatorname{conv} \boldsymbol{S}) = \sup_{\boldsymbol{x}\in\operatorname{conv}\boldsymbol{S}} \psi(\boldsymbol{x}) \leqslant \mu$. □

Lemma A.3 *Let $\mathcal{H}$ be a separable real Hilbert space, let $\varnothing \neq \mathbb{K} \subset \mathbb{N}$, let $E = (e_k)_{k\in\mathbb{K}}$ be an orthonormal basis of $\mathcal{H}$, and let $\alpha \in [0,1]$. For every $k \in \mathbb{K}$, let $\varrho_k \colon \mathbb{R} \to \mathbb{R}$ be α-averaged and such that $\varrho_k(0) = 0$. Define $R \colon \mathcal{H} \to \mathcal{H} \colon x \mapsto \sum_{k\in\mathbb{K}} \varrho_k(\langle x \mid e_k\rangle) e_k$, and fix x and y in $\mathcal{H}$. Then there exists $\Lambda \in \mathscr{D}_{[1-2\alpha,1]}(E)$ such that $Rx - Ry = \Lambda(x-y)$.*

Proof. We saw in Example 3.4 that R is well defined. We have

$$Rx - Ry = \sum_{k\in\mathbb{K}} \big(\varrho_k(\langle x \mid e_k\rangle) - \varrho_k(\langle y \mid e_k\rangle)\big) e_k. \tag{A.2}$$

For every $k \in \mathbb{K}$, there exists a nonexpansive $\theta_k \colon \mathbb{R} \to \mathbb{R}$ such that $\varrho_k = (1-\alpha)\operatorname{Id} + \alpha\theta_k$ and, therefore,

$$\varrho_k(\langle x \mid e_k\rangle) - \varrho_k(\langle y \mid e_k\rangle) = (1-\alpha)(\langle x \mid e_k\rangle - \langle y \mid e_k\rangle) + \alpha\big(\theta_k(\langle x \mid e_k\rangle) - \theta_k(\langle y \mid e_k\rangle)\big). \tag{A.3}$$

Consequently, for every $k \in \mathbb{K}$, there exists $\lambda_k \in [1-2\alpha, 1]$ such that

$$(1-\alpha)\big(\langle x \mid e_k\rangle - \langle y \mid e_k\rangle\big) + \alpha\big(\theta_k(\langle x \mid e_k\rangle) - \theta_k(\langle y \mid e_k\rangle)\big) = \lambda_k(\langle x \mid e_k\rangle - \langle y \mid e_k\rangle). \tag{A.4}$$

We deduce from (A.2) that $Rx - Ry = \sum_{k\in\mathbb{K}} \lambda_k(\langle x \mid e_k\rangle - \langle y \mid e_k\rangle)e_k$, as claimed. □

B Proofs of main results

B.1 Proof of Proposition 3.1

(i): As seen in (1.2), there exists a nonexpansive operator $Q \colon \mathcal{H} \to \mathcal{H}$ such that $R = (1-\alpha)\operatorname{Id} + \alpha Q$. However, by [8, Prop. 4.4 and Cor. 23.9], there exists a maximally monotone operator $A \colon \mathcal{H} \to 2^{\mathcal{H}}$ such that $Q = 2J_A - \operatorname{Id}$. Hence, $R = \operatorname{Id} + \lambda(J_A - \operatorname{Id})$, where $\lambda = 2\alpha \in [0,2]$. For the last claim, notice that, since J_A is firmly nonexpansive [8, Cor. 23.9], so is $R = (1-\lambda)\operatorname{Id} + \lambda J_A$ as a convex combination of two firmly nonexpansive operators [8, Exa. 4.7]. (ii)⇒(i): It follows from [8, Cor. 22.23] that there exists $\phi \in \Gamma_0(\mathbb{R})$ such that $A = \partial\phi$, which provides the expression for R. The increasingness claim follows from Lemma A.1. Finally, if ϕ is even, then prox_ϕ is odd [8, Prop. 24.10] and so is R. (iii): This follows from Lemma A.1.

B.2 Proof of Example 3.3

Let σ_C be the support function of C and set $f = \sigma_C + \phi \circ \|\cdot\| \in \Gamma_0(\mathcal{H})$. Then it follows from [8, Prop. 24.30] and (3.14) that $R = \mathrm{Id} + \lambda(\mathrm{prox}_f - \mathrm{Id})$, However, since prox_f is firmly nonexpansive, it is $1/2$-averaged, which makes R a $\lambda/2$-averaged operator. Now consider the function ϕ of (3.10). Then it is an even function in $\Gamma_0(\mathbb{R})$ with 0 as its unique minimizer. Next, set $\psi = |\cdot| - \arctan|\cdot|$. As seen in Example 3.2(ii), $\phi = \mu\psi^*(\cdot/\mu) - |\cdot|^2/2$ and $\operatorname{dom}\psi^*$ is bounded. Therefore $\operatorname{dom}\phi = \mu\operatorname{dom}\psi^*$ is bounded. In turn, ϕ is supercoercive and we derive from [8, Prop. 14.15] that $\operatorname{dom}\phi^* = \mathbb{R}$. Hence, since $\phi = \phi^{**}$ is strictly convex, it follows derive from [8, Prop. 18.9] that ϕ^* is differentiable on $\mathbb{R}$. In addition, $d_C = \|\cdot\|$. Altogether, (3.14) reduces to

$$(\forall x \in \mathcal{H}) \quad Rx = \begin{cases} \dfrac{\mathrm{prox}_\phi\|x\|}{\|x\|}x, & \text{if } x \neq 0; \\ 0, & \text{if } x = 0 \end{cases} \tag{B.1}$$

and hence, in view of Example 3.2(ii), to (3.15).

B.3 Proof of Example 3.4

Let $x \in \mathcal{H}$ and $y \in \mathcal{H}$. It follows from the nonexpansiveness of the functions $(\varrho_k)_{k\in\mathbb{K}}$ that

$$\sum_{k\in\mathbb{K}} \big|\varrho_k(\langle x \mid e_k\rangle)\big|^2 = \sum_{k\in\mathbb{K}} \big|\varrho_k(\langle x \mid e_k\rangle) - \varrho_k(0)\big|^2 \leqslant \sum_{k\in\mathbb{K}} \big|\langle x \mid e_k\rangle - 0\big|^2 = \|x\|^2. \tag{B.2}$$

Hence, R is well defined. For every $k \in \mathbb{K}$, by (1.2) there exists a nonexpansive function $\theta_k\colon \mathbb{R} \to \mathbb{R}$ such that $\varrho_k = (1-\alpha)\,\mathrm{Id} + \alpha\theta_k$. Hence, $Rx = (1-\alpha)x + \alpha Qx$, where $Qx = \sum_{k\in\mathbb{K}} \theta_k(\langle x \mid e_k\rangle)e_k$. Therefore,

$$\|Qx - Qy\|^2 = \sum_{k\in\mathbb{K}} \big|\theta_k(\langle x \mid e_k\rangle) - \theta_k(\langle y \mid e_k\rangle)\big|^2 \leqslant \sum_{k\in\mathbb{K}} \big|\langle x \mid e_k\rangle - \langle y \mid e_k\rangle\big|^2 = \|x - y\|^2. \tag{B.3}$$

This shows that Q is nonexpansive and hence that R is α-averaged.

B.4 Proof of Example 3.5

Let S be the sorting operator of Example 3.7. Then

$$\begin{aligned} (\forall x \in \mathbb{R}^N)(\forall y \in \mathbb{R}^N) \quad \|Sx - Sy\|^2 &= \|Sx\|^2 - 2\langle Sx \mid Sy\rangle + \|Sy\|^2 \\ &= \|x\|^2 - 2\langle Sx \mid Sy\rangle + \|y\|^2 \\ &\leqslant \|x\|^2 - 2\langle x \mid y\rangle + \|y\|^2 \qquad \text{(B.4)} \\ &= \|x - y\|^2, \qquad \text{(B.5)} \end{aligned}$$

where (B.4) follows from [32, Thm. 368]. This shows that S is nonexpansive. Furthermore, $Q = 2\mathrm{proj}_C - \mathrm{Id}$ is nonexpansive [8, Cor. 4.18]. Note that

$$(1-\omega)\mathrm{proj}_C + \omega S = \left(1 - \frac{1+\omega}{2}\right)\mathrm{Id} + \frac{1+\omega}{2}\left(\frac{1-\omega}{1+\omega}Q + \frac{2\omega}{1+\omega}S\right). \tag{B.6}$$

Since $((1-\omega)Q + 2\omega S)/(1+\omega)$ is nonexpansive as a convex combination of nonexpansive operators, the operator $(1-\omega)\mathrm{proj}_C + \omega S$ is $(1+\omega)/2$-averaged.

B.5 Proof of Example 3.7

Set $A = \mathrm{Diag}(\tau_1, \dots, \tau_{N-1})$. Let x and y be in $\mathbb{R}^{N-1}$, and define $\widetilde{x} = [(Ax)^\top, \theta]^\top$ and $\widetilde{y} = [(Ay)^\top, \theta]^\top$. As seen in (B.5), S is nonexpansive. Consequently,

$$\begin{aligned}\|Rx - Ry\| = \|US\widetilde{x} - US\widetilde{y}\| &\leqslant \|U\|\,\|S\widetilde{x} - S\widetilde{y}\| = \|S\widetilde{x} - S\widetilde{y}\| \\ &\leqslant \|\widetilde{x} - \widetilde{y}\| = \|Ax - Ay\| \leqslant \max\{|\tau_1|, \dots, |\tau_{N-1}|\}\|x - y\|. \end{aligned} \tag{B.7}$$

This shows that R is Lipschitzian with constant $\max\{|\tau_1|, \dots, |\tau_{N-1}|\} < 1$. It is thus α-averaged with $\alpha = (1 + \max\{|\tau_1|, \dots, |\tau_{N-1}|\})/2$ [8, Prop. 4.38].

B.6 Proof of Theorem 4.2

For every $i \in \{1, \dots, m\}$, $P_i = R_i(\cdot + b_i)$ is α_i-averaged and, therefore, there exists a nonexpansive operator $Q_i \colon \mathcal{H}_i \to \mathcal{H}_i$ such that $P_i = (1 - \alpha_i)\,\mathrm{Id} + \alpha_i Q_i$. Since $T = P_m \circ W_m \circ \cdots \circ P_1 \circ W_1$ and P_m is nonexpansive, it suffices to show that

$$\theta_m \text{ is a Lipschitz constant of } W_m \circ \cdots \circ P_1 \circ W_1. \tag{B.8}$$

Let us prove this result by induction. Let $x \in \mathcal{H}_0$ and $y \in \mathcal{H}_0$. If $m = 2$, we derive from the nonexpansiveness of Q_1 that

$$\begin{aligned}&\|(W_2 \circ P_1 \circ W_1)x - (W_2 \circ P_1 \circ W_1)y\| \\ &\quad = \|(W_2 \circ ((1 - \alpha_1)\,\mathrm{Id} + \alpha_1 Q_1) \circ W_1)x - (W_2 \circ ((1 - \alpha_1)\,\mathrm{Id} + \alpha_1 Q_1) \circ W_1)y\| \\ &\quad \leqslant (1 - \alpha_1)\|(W_2 \circ W_1)(x - y)\| + \alpha_1\|(W_2 \circ Q_1 \circ W_1)x - (W_2 \circ Q_1 \circ W_1)y\| \\ &\quad \leqslant (1 - \alpha_1)\|W_2 \circ W_1\|\,\|x - y\| + \alpha_1\|W_2\|\,\|Q_1(W_1 x) - Q_1(W_1 y)\| \\ &\quad \leqslant (1 - \alpha_1)\|W_2 \circ W_1\|\,\|x - y\| + \alpha_1\|W_2\|\,\|W_1(x - y)\| \\ &\quad \leqslant \big((1 - \alpha_1)\|W_2 \circ W_1\| + \alpha_1\|W_2\|\,\|W_1\|\big)\|x - y\|. \end{aligned} \tag{B.9}$$

Hence, T is Lipschitzian with constant

$$(1 - \alpha_1)\|W_2 \circ W_1\| + \alpha_1\|W_2\|\,\|W_1\| = \beta_{2;\varnothing}\|W_2 \circ W_1\| + \beta_{2;\{1\}}\|W_2\|\,\|W_1\| = \theta_2. \tag{B.10}$$

Now assume that $m > 2$ and that (B.8) holds at order $m - 1$. Then

$$\begin{aligned}&\|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)y\| \\ &= \|(W_m \circ ((1 - \alpha_{m-1})\,\mathrm{Id} + \alpha_{m-1} Q_{m-1}) \circ \cdots \circ P_1 \circ W_1)x \\ &\quad - (W_m \circ ((1 - \alpha_{m-1})\,\mathrm{Id} + \alpha_{m-1} Q_{m-1}) \circ \cdots \circ P_1 \circ W_1)y\| \\ &\leqslant (1 - \alpha_{m-1})\|(W_m \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)y\| \\ &\quad + \alpha_{m-1}\|(W_m \circ Q_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ Q_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)y\| \\ &\leqslant (1 - \alpha_{m-1})\|(W_m \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)y\| \\ &\quad + \alpha_{m-1}\|W_m\|\,\|(Q_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (Q_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)y\|. \end{aligned}$$

Hence, the nonexpansiveness of Q_{m-1} yields

$$\begin{aligned}&\|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)y\| \\ &\leqslant (1 - \alpha_{m-1})\|(W_m \circ W_{m-1} \circ P_{m-2} \circ \cdots \circ W_1)x - (W_m \circ W_{m-1} \circ P_{m-2} \circ \cdots \circ W_1)y\| \\ &\quad + \alpha_{m-1}\|W_m\|\,\|(W_{m-1} \circ P_{m-2} \circ \cdots \circ P_1 \circ W_1)x - (W_{m-1} \circ P_{m-2} \circ \cdots \circ P_1 \circ W_1)y\|. \end{aligned} \tag{B.11}$$

On the other hand, the induction hypothesis yields

$$\begin{aligned}
&\|(W_{m-1}\circ P_{m-2}\circ\cdots\circ P_1\circ W_1)x-(W_{m-1}\circ P_{m-2}\circ\cdots\circ P_1\circ W_1)y\|\\
&\leqslant \theta_{m-1}\|x-y\|\\
&=\left(\beta_{m-1;\varnothing}\|W_{m-1}\circ\cdots\circ W_1\|+\sum_{k=1}^{m-2}\sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m-1,k}}\beta_{m-1;\{j_1,\ldots,j_k\}}\sigma_{m-1;\{j_1,\ldots,j_k\}}\right)\|x-y\|.
\end{aligned}\tag{B.12}$$

Similarly, replacing W_{m-1} by $W_m\circ W_{m-1}$ above, we get

$$\begin{aligned}
&\|((W_m\circ W_{m-1})\circ P_{m-2}\circ\cdots\circ P_1\circ W_1)x-((W_m\circ W_{m-1})\circ P_{m-2}\circ\cdots\circ P_1\circ W_1)y\|\\
&\leqslant\left(\beta_{m-1;\varnothing}\|W_m\circ W_{m-1}\circ\cdots\circ W_1\|+\sum_{k=1}^{m-2}\sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m-1,k}}\beta_{m-1;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}}\right)\|x-y\|.
\end{aligned}\tag{B.13}$$

Using (B.11), and then inserting (B.13) and (B.12), we obtain

$$\begin{aligned}
&\|(W_m\circ P_{m-1}\circ\cdots\circ P_1\circ W_1)x-(W_m\circ P_{m-1}\circ\cdots\circ P_1\circ W_1)y\|\\
&\leqslant(1-\alpha_{m-1})\|(W_m\circ W_{m-1}\circ P_{m-2}\circ\cdots\circ W_1)x-(W_m\circ W_{m-1}\circ P_{m-2}\circ\cdots\circ W_1)y\|\\
&\quad+\alpha_{m-1}\|W_m\|\,\|(W_{m-1}\circ P_{m-2}\circ\cdots\circ P_1\circ W_1)x-(W_{m-1}\circ P_{m-2}\circ\cdots\circ P_1\circ W_1)y\|\\
&\leqslant(1-\alpha_{m-1})\times\\
&\quad\left(\beta_{m-1;\varnothing}\|W_m\circ W_{m-1}\circ\cdots\circ W_1\|+\sum_{k=1}^{m-2}\sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m-1,k}}\beta_{m-1;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}}\right)\|x-y\|\\
&\quad+\alpha_{m-1}\|W_m\|\times\\
&\quad\left(\beta_{m-1;\varnothing}\|W_{m-1}\circ\cdots\circ W_1\|+\sum_{k=1}^{m-2}\sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m-1,k}}\beta_{m-1;\{j_1,\ldots,j_k\}}\sigma_{m-1;\{j_1,\ldots,j_k\}}\right)\|x-y\|.
\end{aligned}\tag{B.14}$$

Furthermore, we deduce from (4.3) that

$$(\forall\,\mathbb{J}\subset\{1,\ldots,m-1\})\quad\beta_{m;\mathbb{J}}=\begin{cases}(1-\alpha_{m-1})\beta_{m-1;\mathbb{J}}, & \text{if } m-1\notin\mathbb{J};\\ \alpha_{m-1}\beta_{m-1;\mathbb{J}\smallsetminus\{m-1\}}, & \text{if } m-1\in\mathbb{J}.\end{cases}\tag{B.15}$$

Therefore

$$\begin{cases}\beta_{m;\varnothing}=(1-\alpha_{m-1})\beta_{m-1;\varnothing}\\ \beta_{m;\{j_1,\ldots,j_k\}}=(1-\alpha_{m-1})\beta_{m-1;\{j_1,\ldots,j_k\}} & \text{if } m-1\notin\{j_1,\ldots,j_k\}\\ \beta_{m;\{j_1,\ldots,j_k\}}=\alpha_{m-1}\beta_{m-1;\{j_1,\ldots,j_k\}\smallsetminus\{m-1\}} & \text{if } m-1\in\{j_1,\ldots,j_k\},\end{cases}\tag{B.16}$$

which implies that, if $m-1\notin\{j_1,\ldots,j_k\}$, then $\beta_{m;\{j_1,\ldots,j_k,m-1\}}=\alpha_{m-1}\beta_{m-1;\{j_1,\ldots,j_k\}}$. Hence, (B.14)

yields

$$
\begin{aligned}
&\|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)y\|/\|x-y\| \\
&\quad \leqslant \beta_{m;\varnothing}\|W_m \circ W_{m-1} \circ \cdots \circ W_1\| \\
&\qquad + \sum_{k=1}^{m-2} \sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m-1,k}} (1-\alpha_{m-1})\beta_{m-1;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\qquad + \alpha_{m-1}\beta_{m-1;\varnothing}\|W_m\|\,\|W_{m-1} \circ \cdots \circ W_1\| \\
&\qquad + \sum_{k=1}^{m-2} \sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m-1,k}} \alpha_{m-1}\beta_{m-1;\{j_1,\ldots,j_k\}}\|W_m\|\sigma_{m-1;\{j_1,\ldots,j_k\}} \\
&\quad = \beta_{m;\varnothing}\|W_m \circ W_{m-1} \circ \cdots \circ W_1\| + \beta_{m;m-1}\sigma_{m;\{m-1\}} \\
&\qquad + \sum_{k=1}^{m-2} \sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}\setminus\{m-1\}} \beta_{m;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\qquad + \sum_{k=1}^{m-2} \sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m-1,k}} \beta_{m;\{j_1,\ldots,j_k,m-1\}}\sigma_{m;\{j_1,\ldots,j_k,m-1\}} \\
&\quad = \beta_{m;\varnothing}\|W_m \circ \cdots \circ W_1\| + \sum_{j=1}^{m} \beta_{m;\{j\}}\sigma_{m;\{j\}} \\
&\qquad + \sum_{k=2}^{m-2} \sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}\setminus\{m-1\}} \beta_{m;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\qquad + \sum_{k=2}^{m-1} \sum_{\substack{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k} \\ j_k=m-1}} \beta_{m;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\quad = \beta_{m;\varnothing}\|W_m \circ \cdots \circ W_1\| + \sum_{k=1}^{m-1} \sum_{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}} \beta_{m;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\quad = \theta_m. \qquad \text{(B.17)}
\end{aligned}
$$

Thus, we obtain

$$\|(W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \circ \cdots \circ P_1 \circ W_1)y\| \leqslant \theta_m\|x-y\|, \qquad \text{(B.18)}$$

which establishes (B.8).

B.7 Proof of Proposition 4.3

Define $(\beta_{m;\mathbb{J}})_{\mathbb{J}\subset\{1,\ldots,m-1\}}$ as in (4.3). (i): For every $k \in \{1,\ldots,m-1\}$ and every $(j_1,\ldots,j_k) \in \mathbb{J}_{m,k}$, (4.2) yields

$$\|W_m \circ \cdots \circ W_1\| \leqslant \sigma_{m;\{j_1,\ldots,j_k\}} \leqslant \prod_{i=1}^{m}\|W_i\|. \qquad \text{(B.19)}$$

Consequently, it follows from (4.4) that

$$\|W_m \circ \cdots \circ W_1\| \sum_{\mathbb{J}\subset\{1,\ldots,m-1\}} \beta_{m;\mathbb{J}} \leqslant \theta_m \leqslant \left(\prod_{i=1}^{m}\|W_i\|\right) \sum_{\mathbb{J}\subset\{1,\ldots,m-1\}} \beta_{m;\mathbb{J}}. \qquad \text{(B.20)}$$

In view of (4.3), $(\beta_{m;\mathbb{J}})_{\mathbb{J}\subset\{1,\dots,m-1\}}$ is the discrete probability distribution of a vector of $m-1$ independent Bernoulli random variables. Hence, $\sum_{\mathbb{J}\subset\{1,\dots,m-1\}}\beta_{m;\mathbb{J}}=1$ in (B.20). (ii): For every $i\in\{1,\dots,m-1\}$, $\alpha_i=0$. Therefore, in view of (4.3),

$$(\forall\mathbb{J}\subset\{1,\dots,m-1\})\quad \beta_{m;\mathbb{J}}=\begin{cases}1, & \text{if } \mathbb{J}=\varnothing;\\ 0, & \text{if } \mathbb{J}\neq\varnothing.\end{cases} \tag{B.21}$$

Hence, the result follows from (4.4). (iii): For every $i\in\{1,\dots,m-1\}$, $\alpha_i=1$. Therefore, in view of (4.3),

$$(\forall\mathbb{J}\subset\{1,\dots,m-1\})\quad \beta_{m;\mathbb{J}}=\begin{cases}1, & \text{if } \mathbb{J}=\{1,\dots,m-1\};\\ 0, & \text{if } \mathbb{J}\neq\{1,\dots,m-1\}.\end{cases} \tag{B.22}$$

Invoking (4.4) allows us to conclude. (iv): For every $i\in\{1,\dots,m-1\}$ $\alpha_i=1/2$. Hence, (4.3) yields $(\forall\mathbb{J}\subset\{1,\dots,m-1\})$ $\beta_{m;\mathbb{J}}=2^{1-m}$. Invoking once again (4.4) yields the result. (v): It follows from (4.2) that

$$\begin{aligned}&\sum_{k=1}^{m-1}\sum_{1\leqslant j_1<\ldots<j_k\leqslant m-1}\beta_{m;\{j_1,\dots,j_k\}}\sigma_{m;\{j_1,\dots,j_k\}}\\ &\quad=\sum_{k=1}^{m-1}\sum_{1\leqslant j_1<\ldots<j_k\leqslant m-1}\beta_{m;\{j_1,\dots,j_k\}}\|W_m\circ\cdots\circ W_{j_k+1}\|\,\|W_{j_k}\circ\cdots\circ W_{j_{k-1}+1}\|\\ &\qquad\cdots\|W_{j_1}\circ\cdots\circ W_1\|.\end{aligned} \tag{B.23}$$

We decompose this expression in a sum of terms depending on the value i taken by j_k, namely,

$$\begin{aligned}&\sum_{k=1}^{m-1}\sum_{1\leqslant j_1<\ldots<j_k\leqslant m-1}\beta_{m;\{j_1,\dots,j_k\}}\sigma_{m;\{j_1,\dots,j_k\}}\\ &\quad=\sum_{i=1}^{m-1}\beta_{m;\{i\}}\|W_m\circ\cdots\circ W_{i+1}\|\,\|W_i\circ\cdots W_1\|\\ &\qquad+\sum_{k=2}^{i-1}\sum_{1\leqslant j_1<\ldots<j_{k-1}\leqslant i-1}\beta_{m;\{j_1,\dots,j_{k-1},i\}}\|W_m\circ\cdots\circ W_{i+1}\|\,\|W_i\circ\cdots\circ W_{j_{k-1}+1}\|\\ &\qquad\cdots\|W_{j_1}\circ\cdots\circ W_1\|.\end{aligned} \tag{B.24}$$

In addition, for every $(j_1,\dots,j_{k-1})\in\mathbb{J}_{i,k-1}$, we derive from (4.3) that

$$\begin{aligned}\beta_{m;\{j_1,\dots,j_{k-1},i\}}&=\Bigg(\prod_{j\in\{j_1,\dots,j_{k-1},i\}}\alpha_j\Bigg)\prod_{j\in\{1,\dots,m-1\}\smallsetminus\{j_1,\dots,j_{k-1},i\}}(1-\alpha_j)\\ &=\alpha_i\Bigg(\prod_{j\in\{j_1,\dots,j_{k-1}\}}\alpha_j\Bigg)\Bigg(\prod_{q=i+1}^{m-1}(1-\alpha_q)\Bigg)\prod_{j\in\{1,\dots,i-1\}\smallsetminus\{j_1,\dots,j_{k-1}\}}(1-\alpha_j)\\ &=\alpha_i\Bigg(\prod_{q=i+1}^{m-1}(1-\alpha_q)\Bigg)\beta_{i;\{j_1,\dots,j_{k-1}\}}.\end{aligned} \tag{B.25}$$

Using the above equality in (B.24), factorizing common factors, and invoking (4.4) yields

$$\begin{aligned}
&\sum_{k=1}^{m-1} \sum_{1\leqslant j_1<\ldots<j_k\leqslant m-1} \beta_{m;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\quad = \sum_{i=1}^{m-1} \alpha_i \Bigg(\prod_{q=i+1}^{m-1}(1-\alpha_q)\Bigg)\|W_m\circ\cdots\circ W_{i+1}\|\Big(\beta_{i;\varnothing}\|W_i\circ\cdots W_1\| \\
&\qquad + \sum_{k=2}^{i} \sum_{1\leqslant j_1<\ldots<j_{k-1}\leqslant i-1} \beta_{i;\{j_1,\ldots,j_{k-1}\}}\|W_i\circ\cdots\circ W_{j_{k-1}+1}\|\cdots\|W_{j_1}\circ\cdots\circ W_1\|\Big) \\
&\quad = \sum_{i=1}^{m-1} \alpha_i\theta_i \Bigg(\prod_{q=i+1}^{m-1}(1-\alpha_q)\Bigg)\|W_m\circ\cdots\circ W_{i+1}\|,
\end{aligned} \tag{B.26}$$

and we obtain (4.6).

B.8 Proof of Proposition 4.5

Let $l\in\{1,\ldots,m-1\}$ and set

$$(\forall\,\mathbb{J}\subset\{1,\ldots,m-1\}\smallsetminus\{l\})\quad \beta_{m,l;\mathbb{J}} = \Bigg(\prod_{j\in\mathbb{J}}\alpha_j\Bigg)\prod_{j\in\{1,\ldots,m-1\}\smallsetminus(\mathbb{J}\cup\{l\})}(1-\alpha_j). \tag{B.27}$$

For every $k\in\{1,\ldots,m-1\}$ and every $(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}$, (4.2) yields

$$\sigma_{m;\{j_1,\ldots,j_k\}} \leqslant \sigma_{m;\{j_1,\ldots,j_k\}\cup\{l\}}. \tag{B.28}$$

We infer from (4.4) that

$$\begin{aligned}
&\theta_m(\alpha_1,\ldots,\alpha_{m-1}) \\
&\quad = (1-\alpha_l)\beta_{m,l;\varnothing}\|W_m\circ\cdots\circ W_1\| + \sum_{k=1}^{m-1}\sum_{\substack{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}\\ l\in\{j_1,\ldots,j_k\}}} \alpha_l\beta_{m,l;\{j_1,\ldots,j_k\}\smallsetminus\{l\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\qquad + \sum_{k=1}^{m-2}\sum_{\substack{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}\\ l\notin\{j_1,\ldots,j_k\}}} (1-\alpha_l)\beta_{m,l;\{j_1,\ldots,j_k\}}\sigma_{m;\{j_1,\ldots,j_k\}} \\
&\quad = \beta_{m,l;\varnothing}\big((1-\alpha_l)\|W_m\circ\cdots\circ W_1\| + \alpha_l\|W_m\circ\cdots W_{l+1}\|\,\|W_l\circ\cdots\circ W_1\|\big) \\
&\qquad + \sum_{k=1}^{m-2}\sum_{\substack{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}\\ l\notin\{j_1,\ldots,j_k\}}} \beta_{m,l;\{j_1,\ldots,j_k\}}\big((1-\alpha_l)\sigma_{m;\{j_1,\ldots,j_k\}} + \alpha_l\sigma_{m;\{j_1,\ldots,j_k\}\cup\{l\}}\big).
\end{aligned} \tag{B.29}$$

In view of (B.28) we conclude that

$$\begin{aligned}
\frac{\partial\theta_m}{\partial\alpha_l}(\alpha_1,\ldots,\alpha_{m-1}) &= \beta_{m,l;\varnothing}\big(\|W_m\circ\cdots W_{l+1}\|\,\|W_l\circ\cdots\circ W_1\| - \|W_m\circ\cdots\circ W_1\|\big) \\
&\quad + \sum_{k=1}^{m-2}\sum_{\substack{(j_1,\ldots,j_k)\in\mathbb{J}_{m,k}\\ l\notin\{j_1,\ldots,j_k\}}} \beta_{m,l;\{j_1,\ldots,j_k\}}\big(\sigma_{m;\{j_1,\ldots,j_k\}\cup\{l\}} - \sigma_{m;\{j_1,\ldots,j_k\}}\big) \geqslant 0.
\end{aligned} \tag{B.30}$$

B.9 Proof of Theorem 5.2

(i): For every $i \in \{1,\dots,m\}$, set $P_i = R_i(\cdot + b_i)$ and $(\forall k \in \mathbb{K}_i)\ \pi_{i,k} = \varrho_{i,k}(\cdot + \langle b_i \mid e_{i,k}\rangle)$. Note that, for every $i \in \{1,\dots,m\}$ and every $k \in \mathbb{K}_i$, $\pi_{i,k}$ is α_i-averaged. Furthermore, $(\forall i \in \{1,\dots,m-1\})(\forall x \in \mathcal{H}_i)\ P_i x = \sum_{k\in\mathbb{K}_i} \pi_{i,k}(\langle x \mid e_{i,k}\rangle)e_{i,k}$. Now fix x and y in $\mathcal{H}_0$. It follows from (1.3) and the nonexpansiveness of P_m that

$$\|Tx - Ty\| \leqslant \|(W_m \circ P_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (W_m \circ P_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)y\|. \tag{B.31}$$

In view of Lemma A.3, for every $i \in \{1,\dots,m-1\}$, there exists $\Lambda_i \in \mathscr{D}_{[1-2\alpha_i,1]}(E_i)$ such that

$$\begin{aligned}&(P_i \circ W_i \circ \cdots \circ P_1 \circ W_1)x - (P_i \circ W_i \circ \cdots \circ P_1 \circ W_1)y\\ &\qquad = \Lambda_i\Big((W_i \circ P_{i-1} \circ \cdots \circ P_1 \circ W_1)x - (W_i \circ P_{i-1} \circ \cdots \circ P_1 \circ W_1)y\Big).\end{aligned} \tag{B.32}$$

Recursive application of this identity yields

$$\begin{aligned}&(P_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)x - (P_{m-1} \circ W_{m-1} \circ \cdots \circ P_1 \circ W_1)y\\ &\qquad = (\Lambda_{m-1} \circ W_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)(x-y).\end{aligned} \tag{B.33}$$

This implies that $\|Tx - Ty\| \leqslant \|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|\,\|x-y\|$. Thus,

$$\vartheta_m = \sup_{\substack{\Lambda_1 \in \mathscr{D}_{[1-2\alpha_1,1]}(E_1)\\ \vdots \\ \Lambda_{m-1} \in \mathscr{D}_{[1-2\alpha_{m-1},1]}(E_{m-1})}} \|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|. \tag{B.34}$$

is a Lipschitz constant of T. Set $S = \{1-2\alpha_1,1\}^{\mathbb{K}_1} \times \cdots \times \{1-2\alpha_{m-1},1\}^{\mathbb{K}_{m-1}}$ and $C = [1-2\alpha_1,1]^{\mathbb{K}_1} \times \cdots \times [1-2\alpha_{m-1},1]^{\mathbb{K}_{m-1}}$. For every $i \in \{1,\dots,m-1\}$, $\Lambda_i\colon \mathcal{H}_i \to \mathcal{H}_i$ is generated from a sequence $(\lambda_{i,k})_{k\in\mathbb{K}_i}$ in $[1-2\alpha_i,1]$ via the construction of (5.1). The function

$$\begin{aligned}\psi\colon\ C &\to \mathbb{R}\\ \big((\lambda_{1,k})_{k\in\mathbb{K}_1},\dots,(\lambda_{m-1,k})_{k\in\mathbb{K}_{m-1}}\big) &\mapsto \|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|\end{aligned} \tag{B.35}$$

is convex with respect to each of its coordinates. Hence, we deduce from Lemma A.2 that $\sup \psi(C) = \sup \psi(\operatorname{conv} S) = \sup \psi(S)$, as claimed.

(ii): For every $i \in \{1,\dots,m-1\}$, the identity operator Id_i of $\mathcal{H}_i$ lies in $\mathscr{D}_{\{1-2\alpha_i,1\}}(E_i)$. Hence, $\vartheta_m \geqslant \|W_m \circ \mathrm{Id}_{m-1} \circ \cdots \circ \mathrm{Id}_1 \circ W_1\| = \|W_m \circ \cdots \circ W_1\|$. For every $i \in \{1,\dots,m-1\}$, let $\Lambda_i \in \mathscr{D}_{\{1-2\alpha_i,1\}}(E_i)$ and note that the linear operator

$$\Theta_i = \begin{cases}\dfrac{\Lambda_i - (1-\alpha_i)\,\mathrm{Id}_i}{\alpha_i}, & \text{if } \alpha_i \neq 0;\\ 0, & \text{otherwise}\end{cases} \tag{B.36}$$

is nonexpansive. Using the same kind of decomposition as in the proof of Theorem 4.2 yields

$$\begin{aligned}&\big\|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\big\|\\ &\quad = \big\|W_m \circ \big((1-\alpha_{m-1})\,\mathrm{Id}_{m-1} + \alpha_{m-1}\Theta_{m-1}\big) \circ \cdots \circ \big((1-\alpha_1)\,\mathrm{Id}_1 + \alpha_1\Theta_1\big) \circ W_1\big\| \leqslant \theta_m\end{aligned}$$

and allows us to conclude that $\vartheta_m \leqslant \theta_m$.

B.10 Proof of Proposition 5.5

It follows from (B.34) that

$$\begin{aligned}\vartheta_m(\alpha_1,\dots,\alpha_{m-1}) &= \sup_{\substack{\Lambda_1\in\mathscr{D}_{[1-2\alpha_1,1]}(E_1)\\ \vdots \\ \Lambda_{m-1}\in\mathscr{D}_{[1-2\alpha_{m-1},1]}(E_{m-1})}} \|W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\| \\ &\leqslant \sup_{\substack{\Lambda_1\in\mathscr{D}_{[1-2\alpha'_1,1]}(E_1)\\ \vdots \\ \Lambda_{m-1}\in\mathscr{D}_{[1-2\alpha'_{m-1},1]}(E_{m-1})}} \|W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\| \\ &= \vartheta_m(\alpha'_1,\dots,\alpha'_{m-1}). \end{aligned} \tag{B.37}$$

B.11 Proof of Proposition 5.6

Let us first note that, because of the embeddings, $W_1\colon \mathcal{G}_0\to\mathcal{H}_1$ is continuous and, likewise, every $\Lambda_m\in\mathscr{D}_{[1-2\alpha_m,1]}(E_m)$ is continuous from $\mathcal{H}_m$ to $\mathcal{G}_m$. Hence, for every $(\Lambda_i)_{1\leqslant i\leqslant m}\in\mathscr{D}_{[1-2\alpha_1,1]}(E_1)\times\cdots\times\mathscr{D}_{[1-2\alpha_m,1]}(E_{m-1})$, $\Lambda_m\circ W_m\circ\cdots\circ\Lambda_1\circ W_1\colon\mathcal{G}_0\to\mathcal{G}_m$ is continuous. We now follow the same argument as in the proof of Theorem 5.2. Let x and y be in $\mathcal{G}_0$. For every $i\in\{1,\dots,m\}$, there exists $\Lambda_i\in\mathscr{D}_{[1-2\alpha_i,1]}(E_i)$ such that $Tx-Ty=(\Lambda_m\circ W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1)(x-y)$. Thus, $\|Tx-Ty\|_{\mathcal{G}_m}\leqslant\|\Lambda_m\circ W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}\|x-y\|_{\mathcal{G}_0}$, which leads to (5.6).

B.12 Proof of Corollary 5.7

Since, for every $x\in\mathcal{H}_m$, $(\langle x\mid e_{m,k}\rangle)_{k\in\mathbb{K}_m}\in\ell^2(\mathbb{K}_m)$, it follows from Hölder's inequality that $\|\cdot\|_{\mathcal{G}_m}$ in (5.7) is well defined and does provide a continuous embedding of $\mathcal{H}_m$ in $\mathcal{G}_m$. As in the proof of Theorem 5.2, it is enough to take the supremum in (5.8) over $\boldsymbol{D}=\mathscr{D}_{[1-2\alpha_1,1]}(E_1)\times\cdots\times\mathscr{D}_{[1-2\alpha_{m-1},1]}(E_{m-1})$. For every $i\in\{1,\dots,m\}$, let $\Lambda_i\in\mathscr{D}_{[1-2\alpha_i,1]}(E_i)$. Then

$$\|\Lambda_m\circ W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}\leqslant\|\Lambda_m\|_{\mathcal{G}_m,\mathcal{G}_m}\|W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}. \tag{B.38}$$

Let us designate by $(\lambda_{m,k})_{k\in\mathbb{K}_m}$ the sequence in $[1-2\alpha_m,1]$ involved in the construction of Λ_m in (5.1). If $p<+\infty$, then

$$\begin{aligned}(\forall x\in\mathcal{H}_m)\quad \|\Lambda_m x\|_{\mathcal{G}_m} &= \Bigg\|\sum_{k\in\mathbb{K}_m}\lambda_{m,k}\langle x\mid e_{m,k}\rangle e_{m,k}\Bigg\|_{\mathcal{G}_m} = \Bigg|\sum_{k\in\mathbb{K}_m}\omega_k|\lambda_{m,k}\langle x\mid e_{m,k}\rangle|^p\Bigg|^{1/p} \\ &\leqslant \Bigg|\sum_{k\in\mathbb{K}_m}\omega_k|\langle x\mid e_{m,k}\rangle|^p\Bigg|^{1/p} = \|x\|_{\mathcal{G}_m}, \end{aligned} \tag{B.39}$$

which shows that $\|\Lambda_m\|_{\mathcal{G}_m,\mathcal{G}_m}\leqslant 1$. This inequality holds analogously if $p=+\infty$. We then deduce from (B.38) that $\vartheta_m\leqslant\sup_{(\Lambda_1,\dots,\Lambda_{m-1})\in\boldsymbol{D}}\|W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}$. On the other hand, it follows from (5.6) that

$$\vartheta_m\geqslant\sup_{(\Lambda_1,\dots,\Lambda_{m-1})\in\boldsymbol{D}}\|\operatorname{Id}_m\circ W_m\circ\Lambda_{m-1}\circ\cdots\circ\Lambda_1\circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}, \tag{B.40}$$

which concludes the proof.

B.13 Proof of Proposition 5.10

For every $i \in \{1, \dots, m-1\}$, let $\Lambda_i \in \mathscr{D}_{\{1-2\alpha_i,1\}}(E_i)$ and let $(\lambda_{i,k})_{k\in\mathbb{K}_i}$ be the associated sequence in (5.1). Define

$$(\forall k \in \mathbb{K}_m) \quad \lambda_{m,k} = \begin{cases} -1, & \text{if } (\exists\, (k_0, \dots, k_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1})\ \mu_{k_0,\dots,k_{m-1},k} < 0; \\ 1, & \text{otherwise,} \end{cases} \tag{B.41}$$

and set $\Lambda_m \colon \mathcal{H}_m \to \mathcal{H}_m \colon x \mapsto \sum_{k\in\mathbb{K}_m} \lambda_{m,k}\langle x \mid e_{m,k}\rangle e_{m,k}$ and $V_m = \Lambda_m W_m$. Then, by (5.10),

$$\begin{aligned} &\big(\forall (k_0, \dots, k_m) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_m\big) \\ &\qquad \langle W_1 e_{0,k_0} \mid e_{1,k_1}\rangle \cdots \big\langle W_{m-1} e_{m-2,k_{m-2}} \mid e_{m-1,k_{m-1}}\big\rangle \big\langle V_m e_{m-1,k_{m-1}} \mid e_{m,k_m}\big\rangle \geqslant 0. \end{aligned} \tag{B.42}$$

In addition, it follows from (5.7) and (B.41) that

$$\begin{aligned} \|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} &= \|\Lambda_m \circ V_m \circ \Lambda_{m-1} \circ W_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} \\ &= \|V_m \circ \Lambda_{m-1} \circ W_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}. \end{aligned} \tag{B.43}$$

Therefore, without loss of generality, we assume that

$$\big(\forall (k_0, \dots, k_m) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_m\big) \qquad \mu_{k_0,\dots,k_m} \geqslant 0. \tag{B.44}$$

Let us now show that

$$\|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} \leqslant \|W_m \circ \cdots \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}. \tag{B.45}$$

Let $\varepsilon \in \left]0, +\infty\right[$. Then there exists $x \in \mathcal{H}_0$ such that $\|x\|_{\mathcal{G}_0} = 1$ and

$$\|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} \leqslant \|(W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x\|_{\mathcal{G}_m} + \varepsilon. \tag{B.46}$$

If $p < +\infty$ in (5.7), this yields

$$\begin{aligned} &\|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} \\ &\qquad \leqslant \bigg| \sum_{k_m\in\mathbb{K}_m} \omega_{k_m} \big|\langle (W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m,k_m}\rangle\big|^p \bigg|^{1/p} + \varepsilon. \end{aligned} \tag{B.47}$$

On the other hand,

$$\begin{aligned} &(W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \\ &\qquad = \sum_{k_{m-1}\in\mathbb{K}_{m-1}} \big\langle (\Lambda_{m-1} \circ W_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m-1,k_{m-1}}\big\rangle W_m e_{m-1,k_{m-1}} \end{aligned} \tag{B.48}$$

which, in view of (5.1), implies that

$$\begin{aligned} &(\forall k_m \in \mathbb{K}_m)\ \langle (W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m,k_m}\rangle \\ &\quad = \sum_{k_{m-1}\in\mathbb{K}_{m-1}} \big\langle (\Lambda_{m-1} \circ W_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m-1,k_{m-1}}\big\rangle \big\langle W_m e_{m-1,k_{m-1}} \mid e_{m,k_m}\big\rangle \\ &\quad = \sum_{k_{m-1}\in\mathbb{K}_{m-1}} \lambda_{m-1,k_{m-1}} \big\langle W_m e_{m-1,k_{m-1}} \mid e_{m,k_m}\big\rangle \big\langle (W_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m-1,k_{m-1}}\big\rangle. \end{aligned}$$

Using (5.9) recursively yields

$$(\forall k_m \in \mathbb{K}_m) \quad \langle (W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m,k_m} \rangle = \sum_{(k_0,\ldots,k_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1}} \mu_{k_0,\ldots,k_m} \lambda_{m-1,k_{m-1}} \cdots \lambda_{1,k_1} \langle x \mid e_{0,k_0} \rangle. \tag{B.49}$$

We then deduce from (B.44) that

$$\begin{aligned}(\forall k_m \in \mathbb{K}_m) \quad & \big|\langle (W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m,k_m} \rangle\big| \\ &= \Bigg| \sum_{(k_0,\ldots,k_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1}} \mu_{k_0,\ldots,k_m} \lambda_{m-1,k_{m-1}} \cdots \lambda_{1,k_1} \langle x \mid e_{0,k_0} \rangle \Bigg| \\ &\leqslant \sum_{(k_0,\ldots,k_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1}} \mu_{k_0,\ldots,k_m} |\lambda_{m-1,k_{m-1}}| \cdots |\lambda_{1,k_1}| \, |\langle x \mid e_{0,k_0} \rangle| \\ &\leqslant \sum_{(k_0,\ldots,k_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1}} \mu_{k_0,\ldots,k_m} \, |\langle x \mid e_{0,k_0} \rangle| \,. \end{aligned} \tag{B.50}$$

Set $y = \sum_{k_0 \in \mathbb{K}_0} |\langle x \mid e_{0,k_0} \rangle| \, e_{0,k_0}$. In view of (5.12), $\|y\|_{\mathcal{G}_0} = \|x\|_{\mathcal{G}_0} = 1$. Thus, (B.50) yields

$$\begin{aligned}\big|\langle (W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1)x \mid e_{m,k_m} \rangle\big| &\leqslant \sum_{(k_0,\ldots,k_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1}} \mu_{k_0,\ldots,k_m} \langle y \mid e_{0,k_0} \rangle \\ &= \langle (W_m \circ \cdots \circ W_1)y \mid e_{m,k_m} \rangle. \end{aligned} \tag{B.51}$$

It then follows from (B.46) and the fact that $\|y\|_{\mathcal{G}_0} = 1$ that

$$\begin{aligned}\|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} &\leqslant \Bigg| \sum_{k_m \in \mathbb{K}_m} \omega_{k_m} |\langle (W_m \circ \cdots \circ W_1)y \mid e_{m,k_m} \rangle|^p \Bigg|^{1/p} + \varepsilon \\ &\leqslant \|(W_m \circ \cdots \circ W_1)y\|_{\mathcal{G}_m} + \varepsilon \\ &\leqslant \|W_m \circ \cdots \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} + \varepsilon. \end{aligned} \tag{B.52}$$

The same inequality is obtained similarly for $p = +\infty$. This establishes (B.45), which leads to

$$\sup_{\substack{\Lambda_1 \in \mathscr{D}_{\{1-2\alpha_1,1\}}(E_1) \\ \vdots \\ \Lambda_{m-1} \in \mathscr{D}_{\{1-2\alpha_{m-1},1\}}(E_{m-1})}} \|W_m \circ \Lambda_{m-1} \circ \cdots \circ \Lambda_1 \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m} \leqslant \|W_m \circ \cdots \circ W_1\|_{\mathcal{G}_0,\mathcal{G}_m}. \tag{B.53}$$

Since the converse inequality holds straightforwardly, the proof is complete.

B.14 Proof of Proposition 5.12

We use arguments similar to those of the proof of Proposition 5.10. For every $i \in \{1,\ldots,m-1\}$, let $\Lambda_i \in \mathscr{D}_{\{1-2\alpha_i,1\}}(E_i)$. There exists $x \in \mathcal{H}_0$ such that $\|x\|_{\mathcal{G}_0} = 1$ and

$$\|W_m \Lambda_{m-1} \cdots \Lambda_1 W_1\|_{\mathcal{G}_0,\mathcal{G}_m} = \|(W_m \Lambda_{m-1} \cdots \Lambda_1 W_1)x\|_{\mathcal{G}_m}. \tag{B.54}$$

On the other hand, for every $k_m \in \mathbb{K}_m$,

$$\big|\langle W_m \Lambda_{m-1} \cdots \Lambda_1 W_1 x \mid e_{m,k_m} \rangle\big| \leqslant \sum_{(k_0,\ldots,k_{m-1}) \in \mathbb{K}_0 \times \cdots \times \mathbb{K}_{m-1}} |\mu_{k_0,\ldots,k_m}| \, |\langle x \mid e_{0,k_0} \rangle| \,. \tag{B.55}$$

Setting $y = \sum_{k_0 \in \mathbb{K}_0} |\langle x \mid e_{0,k_0} \rangle| \, e_{0,k_0}$ yields $\big|\langle W_m \Lambda_{m-1} \cdots \Lambda_1 W_1 x \mid e_{m,k_m} \rangle\big| \leqslant \langle (A_m \cdots A_1)y \mid e_{m,k_m} \rangle$, and (B.54) implies that $\|W_m \Lambda_{m-1} \cdots \Lambda_1 W_1\|_{\mathcal{G}_0,\mathcal{G}_m} \leqslant \|A_m \cdots A_1 y\|_{\mathcal{G}_m} \leqslant \|A_m \cdots A_1\|_{\mathcal{G}_0,\mathcal{G}_m}$, which concludes the proof.

References

[1] N. Akhtar and A. Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[2] C. H. Aladag, E. Egrioglu, and U. Yolcu, Robust multilayer neural network based on median neuron model, *Neural Comput. Appl.*, vol. 24, pp. 945–956, 2014.

[3] A. Athalye, N. Carlini, and D. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, *Proc. Intl. Conf. Machine Learn.*, pp. 274–283, 2018.

[4] J.-B. Baillon, R. E. Bruck, and S. Reich, On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces, *Houston J. Math.*, vol. 4, pp. 1–9, 1978.

[5] R. Balan, M. Singh, and D. Zou, Lipschitz properties for deep convolutional networks, 2017. https://arxiv.org/abs/1701.05217.pdf

[6] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, Spectrally-normalized margin bounds for neural networks, *Adv. Neural Inform. Process. Syst.*, vol. 30, pp. 6240–6249, 2017.

[7] M. Basirat and P. M. Roth, The quest for the golden activation function, arxiv, 2018. https://arxiv.org/pdf/1808.00783

[8] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces,* 2nd ed., corrected reprint. Springer, New York, 2019.

[9] I. Bayram, On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty, *IEEE Trans. Signal Process.*, vol. 64, pp. 1597–1608, 2016.

[10] C. Bertocchi, E. Chouzenoux, M.-C. Corbineau, J.-C. Pesquet, and M. Prato, Deep unfolding of a proximal interior point method for image restoration, *Inverse Problems*, vol. 36, art. 034005, 2020.

[11] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, Optimal approximation with sparsely connected deep neural networks, *SIAM J. Math. Data Sci.*, vol. 1, pp. 8–45, 2019.

[12] J. M. Borwein, G. Li, and M. K. Tam, Convergence rate analysis for averaged fixed point iterations in common fixed point problems, *SIAM J. Optim.*, vol. 27, pp. 1–33, 2017.

[13] R. I. Boţ and E. R. Csetnek, A dynamical system associated with the fixed points set of a nonexpansive operator, *J. Dynam. Differential Equations,* vol. 29, pp. 155–168, 2017.

[14] Y.-L. Boureau, J. Ponce, and Y. LeCun, A theoretical analysis of feature pooling in visual recognition, *Intl. Conf. Machine Learn.*, pp. 111–118, 2010.

[15] M. Bravo and R. Cominetti, Sharp convergence rates for averaged nonexpansive maps, *Israel J. Math.*, vol. 227, pp. 163–188, 2018.

[16] L. M. Briceño-Arias and P. L. Combettes, Monotone operator methods for Nash equilibria in non-potential games, in *Computational and Analytical Mathematics,* (D. Bailey *et. al.*, eds.), pp. 143–159. Springer, New York, 2013.

[17] J. Bruna and S. Mallat, Invariant scattering convolution networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 1872–1886, 2013.

[18] N. Carlini and D. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, *Proc. ACM Workshop Artificial Intell. Security,* pp. 3–14, 2017.

[19] J. Chorowski and J. M. Zurad, Learning understandable neural networks, with nonnegative weight constraints, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, pp. 62–69, 2015.

[20] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, Fast and accurate deep network learning by exponential linear units (ELUs), arxiv, 2015. https://arxiv.org/abs/1511.07289

[21] P. L. Combettes, Solving monotone inclusions via compositions of nonexpansive averaged operators, *Optimization,* vol. 53, pp. 475–504, 2004.

[22] P. L. Combettes and L. E. Glaudin, Quasinonexpansive iterations on the affine hull of orbits: From Mann's mean value algorithm to inertial methods, *SIAM J. Optim.*, vol. 27, pp. 2356–2380, 2017.

[23] P. L. Combettes and J.-C. Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases, *SIAM J. Optim.*, vol. 18, pp. 1351–1376, 2007.

[24] P. L. Combettes and J.-C. Pesquet, Deep neural network structures solving variational inequalities, *Set-Valued Var. Anal.*, published online 2020-02-13.

[25] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.*, vol. 4, pp. 1168–1200, 2005.

[26] L. Condat, A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms, *J. Optim. Theory Appl.*, vol. 158, pp. 460–479, 2013.

[27] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer, New York, 2003.

[28] S. Geman and D. E. McClure, Bayesian image analysis: An application to single photon emission tomography, *Proc. Statist. Comput. Section Amer. Stat. Assoc.*, pp. 12–18, 1985.

[29] X. Glorot, A. Bordes, and Y. Bengio, Deep sparse rectifier neural networks, *Proc. 14th Int. Conf. Artificial Intell. Stat.*, pp. 315–323, 2011.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, arxiv, 2014. https://arxiv.org/abs/1412.6572

[31] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, Maxout networks, *Proc. 30th Intl. Conf. Machine Learn.*, pp. 1319–1327, 2013.

[32] G. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities,* 2nd ed. Cambridge Univ. Press, Cambridge, 1952.

[33] M. Hein and M. Andriushchenko, Formal guarantees on the robustness of a classifier against adversarial manipulation, *Adv. Neural Inform. Process. Syst.*, vol. 30, pp. 2266–2276, 2017.

[34] S. Ko, D. Yu, and J.-H. Won, Easily parallelizable and distributable class of algorithms for structured sparsity, with optimal acceleration, *J. Comput. Graph. Stat.*, to appear.

[35] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, Fooling end-to-end speaker verification with adversarial examples, *Proc. IEEE Intl. Conf. Acoustic, Speech Signal Process.*, pp. 1962–1966, 2018.

[36] A. Krizhevsky, Convolutional deep belief networks on CIFAR-10, technical report, University of Toronto, 2010. https://www.cs.toronto.edu/~kriz/conv-cifar10-aug2010.pdf

[37] C.-Y. Lee, P. W. Gallagher, and Z. Tu, Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree, *Proc. Machine Learn. Res.*, vol. 51, pp. 464–472, 2016.

[38] A. L. Maas, A. Y. Hannun, and A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, *Proc. 30th Int. Conf. Machine Learn.*, 2013.

[39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, arxiv, 2017. https://arxiv.org/abs/1706.06083

[40] H. N. Mhaskar and C. A. Micchelli, How to choose an activation function, *Adv. Neural Inform. Process. Syst.*, pp. 319–326, 1994.

[41] W. M. Moorsi, The forward-backward algorithm and the normal problem, *J. Optim. Theory Appl.*, vol. 176, pp. 605–624, 2018.

[42] M. Nakagawa, An artificial neuron model with a periodic activation function, *J. Phys. Soc. Japan,* vol. 64, pp. 1023–1031, 1995.

[43] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, *Proc. IEEE Symp. Security Privacy*, pp. 582–597, 2016.

[44] A. Raghunathan, J. Steinhardt, and P. Liang, Certified defenses against adversarial examples, *Proc. Intl. Conf. Learn. Represent.*, 2018. https://arxiv.org/pdf/1801.09344.pdf

[45] P. Ramachandran, B. Zoph, and Q. V. Le, Searching for activation functions, *Proc. Intl. Conf. Learn. Represent.*, 2018. https://arxiv.org/pdf/1710.05941.pdf

[46] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Rev.*, vol. 65, pp. 386–408, 1958.

[47] W. Ruan, X. Huang, and M. Kwiatkowska, Reachability analysis of deep neural networks with provable guarantees, *Proc. 27th Intl. Joint Conf. Artificial Intell.*, pp. 2651–2659, 2018.

[48] S. Sabour, N. Frosst, and G. E. Hinton, Dynamic routing between capsules, *Adv. Neural Inform. Process. Syst.*, vol. 30, pp. 3856–3866, 2017.

[49] K. Scaman and A. Virmaux, Lipschitz regularity of deep neural networks: Analysis and efficient estimation, *Adv. Neural Inform. Process. Syst.*, vol. 31, pp. 3839–3848, 2018.

[50] J. Sokolić, R. Giryes, G. Sapiro, and M. R. D. Rodrigues, Robust large margin deep neural networks, *IEEE Trans. Signal Process.*, vol. 65, pp. 4265–4280, 2017.

[51] Y. Sun, B. Wohlberg, and U. S. Kamilov, An online plug-and-play algorithm for regularized image reconstruction, *IEEE Trans. Comput. Imaging,* vol. 5, pp. 395–408, 2019.

[52] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, Intriguing properties of neural networks, arxiv, 2013. https://arxiv.org/pdf/1312.6199.pdf

[53] Y. Tsuzuku, I. Sato, and M. Sugiyama, Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks, *Adv. Neural Inform. Process. Syst.*, vol. 31, pp. 6541–6550, 2018.

[54] E. Wong and J. Z. Kolter, Provable defenses against adversarial examples via the convex outer adversarial polytope, *Proc. 35th Int. Conf. Machine Learn.*, vol. 80, pp. 5286–5295, 2018.

[55] M. Yamagishi and I. Yamada, Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization, *Inverse Problems,* vol. 33, art. 044003, 35 pp., 2017.

[56] P. Yi and L. Pavel, Distributed generalized Nash equilibria computation of monotone games via double-layer preconditioned proximal-point algorithms, *IEEE Trans. Control Network Syst.*, vol. 6, pp. 299–311, 2019.

[57] E. Zeidler, *Nonlinear Functional Analysis and Its Applications.* Springer, New York, 1985–1990.

[58] Q. Zhao and L. D. Griffin, Suppressing the unusual: Towards robust CNNs using symmetric activation functions, arxiv, 2016. https://arxiv.org/abs/1603.05145.pdf